KU LEUVEN

FACULTY OF SCIENCE

Subjects, Trials, and Levels: Statistical Power in Conjoint Experiments

Alberto Stefanelli

Supervisor: Prof. Patrick Onghena
Affiliation *KU Leuven*

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics and Data Science

Academic year 2022-2023

Preface

Power considerations are often disregarded in the design of conjoint experiments, despite the importance of power analysis for survey-experimental techniques. This thesis provides a general framework to calculate power in fully randomised conjoint experiments, a type of factorial experimental design popularised in the last years in political science. I use information from an extensive literature review of conjoint experiments in top political science journals to set up a simulation. The simulation uses a flexible data-generating model that allows the exploration of the statistical properties of a wide range of commonly employed designs and provides guidance on how various design choices impact the statistical power of conjoint experiments.

I would like to acknowledge the support and guidance of my supervisor, Patrick Onghena, for his valuable feedback and suggestions. I am also grateful to Mirjam Moerbeek, Fredrik Sävje, Juraj Medzihorsky, Constantin Manuel Bosancianu, Bruno Castanho Silva, and Levente (Levi) Littvay for their support and the constructive criticism of some parts of the manuscript. I would also like to thank the participants of the course *Introduction to Conjoint Experiments* that I thought during the past few years for encouraging me to learn more about conjoint experiments.

I hope that the discussion and the results presented in this thesis allow researchers to understand the importance of power analysis for conjoint experiments and, in turn, to achieve adequate design for future studies.

Contribution statement

The idea for this thesis crystallised during the Master of Science in Statistics and Data Science Program (Quantitative Analysis in the Social Sciences) at KU Leuven. Because of the importance of power analysis in conjoint experiments for applied researchers, a working paper building on the idea of this thesis was written (together with Martin Lukac) and published on the Socarxiv platform. Below, I clarify my personal contribution to the thesis:

- Original idea of the thesis topic: the broad idea of the thesis has been formulated by myself. I am also responsible for situating the thesis within the methodological literature in political science.
- The literature review: I am solely responsible for the literature review in this thesis. It includes a brief history of conjoint experiments, the increased popularity of conjoint experiments in political science, the discussion on Hainmueller, Hopkins, and Yamamoto (2014) framework, and a section on power analysis.
- Power analysis in conjoint experiments: the initial write-up of the data generating process and the model for the conjoint experiment has been designed together with Martin Lukac and further elaborated by myself in subsequent versions of the manuscript and this thesis.
- The simulation setup: the literature review of conjoint experiments in political science has been carried out entirely by myself. This includes the coding of the conjoint coefficients, sample size, tasks, attributes, and levels of the sampled conjoint experiments. The simulation write-up has been written together with Martin Lukac and further elaborated by myself in subsequent versions of the manuscript and this thesis.
- Results: the results have been written up by myself with minor editing help from Martin Lukac. This included the evaluation of simulation results, the trade-offs between power and the various design aspects, and the retrospective power analysis.
- Conclusions: the initial write-up of the conclusion has been written together with Martin Lukac. The limitations of the study have been written entirely by myself.
- R code: the R code for the simulation and the analysis has been written together with Martin Lukac. The code for the web application has been written by Martin Lukac and further developed by myself to include additional features and bug fixes.

Summary

Conjoint experiments have become extremely popular in political science in the past years. The seminal paper of Hainmueller, Hopkins, and Yamamoto (2014) has contributed to popularising the use of conjoint experiments to answer multi-dimensional research questions with the presence of multiple experimental attributes that vary at the same time. Despite the importance of power analysis for survey-experimental techniques, researchers often disregard power calculations in the context of conjoint experiments. The main goal of this thesis is to show that conjoint designs, like any other factorial design, are subjected to rather strict power requirements. Specifically, I aim to show how the number of experimental attributes, trials, and subjects impact the statistical power of conjoint experiments.

To this end, I begin by tracing the history of conjoint experiments, contextualising the thesis within the contemporary empirical and methodological scholarship in political science. Following, I provide a brief overview of statistical power and underscore the potential threats to validity, reliability, and replicability that can arise from neglecting power considerations. Then, I introduce a simulation-based framework that leverages a flexible data-generating model to simulate conjoint data. Subsequently, I conduct an extensive literature review to gather information on the experimental designs typically employed in conjoint studies in political science. Using this information, I run a simulation to explore the statistical properties of a wide range of commonly employed conjoint designs. Results show that—even with relatively large sample size and numbers of trials—conjoint experiments are not well suited to draw inferences for designs with large numbers of experimental attributes (≥ 15) and relatively small effect sizes (≤ 0.05).

To gauge the extent to which the experimental designs typically employed in political science diverge from conventional power recommendations ($\alpha = .80$), I conducted a retrospective power analysis using the information obtained from the literature review. Results show that about one-third of the studies that I examined do not have sufficient power to discover the reported effects and are more prone to errors of sign (Type S error) and magnitude (Type M error) compared to designs recruiting larger samples or asking respondents for more trials. This supports the idea that disregarding power considerations for conjoint experimental designs can have serious consequences for theory testing and, more importantly, for the validity of scientific results. Hence, to help researchers to achieve adequate designs for future conjoint experiments, I develop a web application that can be used to perform *a priori* power analysis. This web application is accessible from any browser and is usable by researchers without any statistical background in conjoint experiments and statistical power. I conclude the thesis by pointing out the limitations of the used approach and a few recommendations for further research on the topic.

Acronyms

- AMCE: Average Marginal Component Effect
- Type S error: Error of sign
- Type M error: Error of magnitude

List of Figures

Tables

Contents

1 Introduction

> "Since statistical significance is so earnestly sought and devoutly wished for by behavioural scientists, one would think that the *a priori* probability of its accomplishment would be routinely determined and well understood."
> — Jacob Cohen 1962

Research in the field of political science and public policy has long been interested in why individuals choose one candidate, party, or policy over another. One of the main characteristics of such decision-making processes is that the available options vary on multiple dimensions. While traditional survey experiments remain constrained to examining only one or two factors (Gaines, Kuklinski, and Quirk 2007), factorial experiments allow the investigation of multiple experimental characteristics in a single design. In the past few years, one type of factorial experiment that has started to be extensively used in political science and, more in general, in the social and behavioural sciences is the conjoint design proposed by Hainmueller, Hopkins, and Yamamoto (2014).

Despite the increasing popularity of such design, researchers often disregard one of the most fundamental tools for designing conjoint studies: power analysis (Cohen 1962). Power analysis consists of assessing the probability of successfully rejecting the null hypothesis when it is *false*. In other words, it can be used to estimate the required sample size to detect an effect of a given size with a certain degree of confidence and precision. Power analysis also reduces the chances of obtaining false positives or exaggerated findings (Gelman and Carlin 2014). The difficulty in replicating some of the experimental studies or the emergence of contradictory results in the literature are caused—*inter alia*—by underpowered experimental designs (Open Science Collaboration 2015; Maxwell 2004). The reason for it is that studies with low power are more likely to yield large and significant effects that are due to random fluctuations alone. Finally, (*a priori*) power analysis can save costs by cautioning against overpowered designs with a much higher number of respondents (and, thus, higher costs) than necessary for discovering the hypothesised effect.

Since the publication of the seminal work by Hainmueller, Hopkins, and Yamamoto (2014), the assumption that conjoint experiments are less sensitive to power constraints has become a common belief among researchers. Scholarship published in top journals in political science has argued that conjoint designs "free us from the power constraints that limit traditional factorial experiments" (Joshua D Kertzer, Renshon, and Yarhi-Milo 2019, p. 7) and solve key problems in experimental research such as "the trade-off between statistical power and the desire to employ many experimental conditions" (Knudsen and Johannesson 2018, p. 2). This is further complicated by the fact that calculating power for a conjoint experiment is not a trivial exercise. Conjoint experiments typically involve the use of multiple treatments, repeated tasks, and paired vignettes which add a substantial layer of complexity to the analysis of statistical power. Estimating the required statistical power by taking into account the (a) trade-off between the number of experimental attributes, (b) the number of respondents and tasks that each respondent shall undertake, and (c) a range of hypothesised effect size is a complex task that requires extensive and technical knowledge of both statistical power and conjoint literature.

This thesis contributes to the literature on conjoint experiments by developing a general framework for calculating power for conjoint experiments using simulation techniques (Arnold et al. 2011; Astivia, Gadermann, and Guhn 2019). Specifically, it provides guidance concerning (a) the required sample size—given as a combination of the number of respondents and the number of tasks performed by each respondent—and (b) the maximum number of experimental conditions—i.e., the levels—used in the experiment. To achieve this purpose, I first conducted an extensive literature review to gather information on which experimental designs are typically employed in conjoint studies and the plausible effect sizes discovered in the literature. I restrict the literature review to the 14 most cited academic journals in political science and neighbouring disciplines. For each article included in the literature review, I extracted (a) the estimated effect size, and (b) their statistical significance. In addition, I collected information on the design choices made by the authors in regard to the number of (c) respondents, (d) tasks undertaken

by each respondent, and (e) experimental attributes (i.e., levels) included in the design. The resulting data set is used to derive the range of possible values that each design attribute usually takes and subsequently used to set up a large simulation study. The data-generating mechanism for the simulation is designed to be as flexible as possible using insight from previous research on how respondents typically answer forced-choice conjoint experiments (see, Jenke et al. 2020; Meißner and Decker 2010).

The results show that, even with relatively large samples and numbers of trials, conjoint experiments are not suited to draw inferences for experiments that discover relatively small effect sizes and a large number of experimental conditions. Specifically, the precision of the estimated effects rapidly decreases for designs with a relatively small number of respondents sizes (≤ 1000) and trials (≤ 3) or with a high number of levels (≥ 15) that find small but statistically significant effects (≤ 0.03). Next, to gauge the extent to which the experimental designs typically employed in political science diverge from conventional power recommendations ($\alpha \geq .80$) (Cohen 1962), I conducted a retrospective power analysis for each of the articles included in the literature review. Results show that about one-third of the studies that I examined do not have sufficient power to discover some of the reported effects and are more prone to errors of sign and magnitude (Gelman and Carlin 2014) compared to designs recruiting larger samples or asking respondents for more trials. This supports the idea that disregarding power considerations for conjoint designs can have serious consequences for theory testing and could lead to biased results that are unlikely to replicate in subsequent studies.

A web application has been developed to aid the design of conjoint studies. The application can be used to perform sample size calculations for a variety of popular conjoint designs using the results of the simulation. It is built using the Shiny package (W. Chang et al. 2017) in the R programming language (R Core Team 2019); can be accessed on any device with an internet browser; and requires no programming knowledge to use Lukac and Stefanelli (2020). In addition, I made the R code to simulate and calculate power public[1].

[1]The R code to replicate the results of this thesis and to calculate the power is contained on the author's GitHub and can be found at this link.

The code can be easily adapted to more complex conjoint designs with a larger number of tasks (≥ 15), treatment effect heterogeneity across different segments of the population (e.g. multiple groups), within-subject conjoint designs, and/or less common sampling procedures (e.g. two-stage sampling). This makes the simulation-based approach proposed in this work a more flexible framework compared to the usage of non-parametric techniques proposed in some of the work that appeared after the working paper of Stefanelli and Lukac (2020) (which is based on the work presented in the thesis) such as Schuessler and Freitag (2020).

The thesis is organised as follows. First, I introduce the fundamental features of conjoint designs and their history, detail the characteristics of the fully randomised design proposed by Hainmueller, Hopkins, and Yamamoto (2014), and review the statistical properties of the most commonly used causal quantity of interest that is used to analyse conjoint data, the Average Marginal Component Effects (AMCE). Next, I describe the rationale of power analysis, its importance in the context of experimental techniques, and its relevance for conjoint designs. I also introduce two other types of design measures related to power analysis, namely Type S (sign) and Type M (magnitude) error rates (Gelman and Carlin 2014). Next, I discuss the conceptualisation of the assumed data-generating process. I, then, lay out the methodology and the findings of an extensive literature review that summarises the most common conjoint designs used in the social and behavioural sciences. Using this information, I setup up a series of simulations that explore the statistical properties of the AMCE (Hainmueller, Hopkins, and Yamamoto 2014).

Subsequently, I outline the results of the simulation, focusing on the conventional measure of power analysis (Cohen 1962) and design measures of Type S (sign) and Type M (magnitude) error rates (Gelman and Carlin 2014). I interpret the findings in light of the current conjoint literature. I do so by performing a *post hoc* power analysis for the studies included in our literature review using a wide range of assumed population effects. Finally, I introduce the web application and detail the procedure to calculate statistical

power and design error rates using the designed software and the R code. I conclude by highlighting the limitations of this study and I provide a few suggestions for future research on power analysis in conjoint experiments.

2 Literature Review

2.1 A brief history of conjoint experiments

Conjoint analysis– a survey-experimental technique of widespread use across various domains within the social sciences– is a tool to evaluate multi-dimensional preferences. Conjoint methodologies have their roots in the statistical literature on the design of experiments, with seminal works such as Cox's seminal contribution in 1958. Cox introduced a comprehensive experimental framework to systematically manipulate and control variables and assess their impact on an outcome of interest. These theoretical insights, originally conceived for applications in the industrial and agricultural sectors, lay the foundations for the development of factorial experiments. In general terms, factorial designs are a class of experimental techniques that involve varying two or more factors (i.e., experimental conditions or treatment) simultaneously to observe their individual and combined effects on a response variable.

Conjoint analysis borrows from the conceptual framework of factorial experiments developed by Cox and applies their main insights to human decision-making and preference formation. Developed in the field of mathematical psychology during the 1960s, this type of design takes the name of "conjoint measurement" due to its goal of assessing the "conjoined" or "combined" effects of various factors on individuals' preferences. The methodological framework at the base of conjoint experiments was proposed and popularised by R. Duncan Luce 1964 and Amos Tversky 1967. Their aim: deriving axiomatic theories that would provide insights into the factors upon which complex human decisions are made. These factors, often representing different attributes, characteristics, or dimensions of a choice or scenario, could then be studied individually or in combination to better understand how they contribute to overall decision outcomes. In doing so, this class of methods aimed to estimate the relative and combined importance and influence of specific attributes on the overall (stated) preferences of individuals.

Conjoint experiments are considered able to "reverse-engineer" the decision-making

process that governs an individual's choices. They probe how individuals allocate importance to different attributes and attribute combinations. For that reason, conjoint experiments started to be widely adopted in marketing and neighbouring disciplines. The seminal work of Green and Rao (1971) employed the name "conjoint analysis" to describe a cluster of survey-experimental approaches aimed at analysing consumer preferences and choices within the fields of marketing, product development, and business. Their contribution brought into the limelight conjoint analysis to evaluate two or more alternative profiles (e.g., products) with varying attributes, typically presented in a tabular format. Their work also popularised the usage of conjoint analysis to assess the potential trade-off between different attributes and attribute levels.

For instance, researchers may be interested in what features of a smartphone drive customers' purchases. These aspects encompass factors like the size of the screen, the duration of the battery life, the quality of the camera, and the price of the device. Each of these attributes has distinct levels (e.g., low, medium, and high camera quality). In the case of a force-choice conjoint experiment, respondents are presented with two hypothetical smartphones that differ randomly in terms of attribute levels. Respondents are then asked to choose one of the two smartphones based on the provided information. In this way, researchers can understand the relative and combined importance of each of these factors. Moreover, they can also evaluate which features participants are inclined to give up (i.e., trade-off) in exchange for specific other attributes. For instance, whether respondents are willing to buy a smartphone with lower camera quality when the price of the smartphone surpasses the average market price.

Conducted under the name of "stated choice methods" or "stated-preference analysis", conjoint designs have found substantial utilisation not only within the domain of marketing but also in the field of economics. For instance, Wiktor Adamowicz et al. (1998) use conjoint experiments to explore citizens' preferences towards specific environmental policies and regulations. The authors apply conjoint techniques to estimate the economic value that individuals place on non-market goods or services, such as clean air, water

quality, or biodiversity (for other examples of conjoint application on the same topic see, W. Adamowicz, J. Louviere, and Williams 1994; Boxall et al. 1996; Mackenzie 1993). Another prominent usage of conjoint experiments is in health economics and related disciplines (e.g., Ryan and Farrar 2000; F. R. Johnson, Banzhaf, and Desvousges 2000). In this context, conjoint experiments have been employed to understand individuals' preferences, trade-offs, and willingness to pay for different types of healthcare services and improvements in health and life expectancy, which have significant implications for health policy and economic analysis.

2.2 The popularity of conjoint experiment in political science

Although not as popular as in the fields of marketing and econometrics, conjoint experiments have also been used to investigate societal and political phenomena. Jasso and Rossi's work 1977 has used conjoint methodologies to explore social stratification and individuals' perception of fairness. Specifically, the researchers were interested in how individuals make judgments about fairness and distributive justice in the context of salary allocations (for a review of conjoint experiments in sociology see, Wallander 2009). Another example is Loewen, Rubenson, and Spirling (2012). Using discrete-choice conjoint experiments, the authors explore the factors that underlie the endorsement of electoral reforms in the realm of voting behaviour (i.e., referendum) (Loewen, Rubenson, and Spirling 2012).

In recent years, there has been a noticeable surge in the utilisation of conjoint designs in the social sciences and, more specifically, in political science (see Figure 1). Due to their versatility, political scientists have started to employ conjoint analysis across a wide spectrum of topics. These include inquiries from how citizens define terrorism (e.g., Huff and Joshua D. Kertzer 2018) to public sentiments regarding immigrants (e.g., Ward 2019), understanding the factors that shape perceptions of corruption severity (e.g., Martin 2019), mapping citizens' belief systems (e.g., Goggin, Henderson, and Theodoridis 2019), and delving into the trade-offs that individuals are willing to make when choosing between

political candidates or parties (e.g., Franchino and Zucchini 2015). The popularity of conjoint analysis extends even to political news media: in the spring of 2019, CBS News featured a conjoint experiment on television that assessed the attributes that US voters value the most when selecting presidential candidates (Khanna 2019).

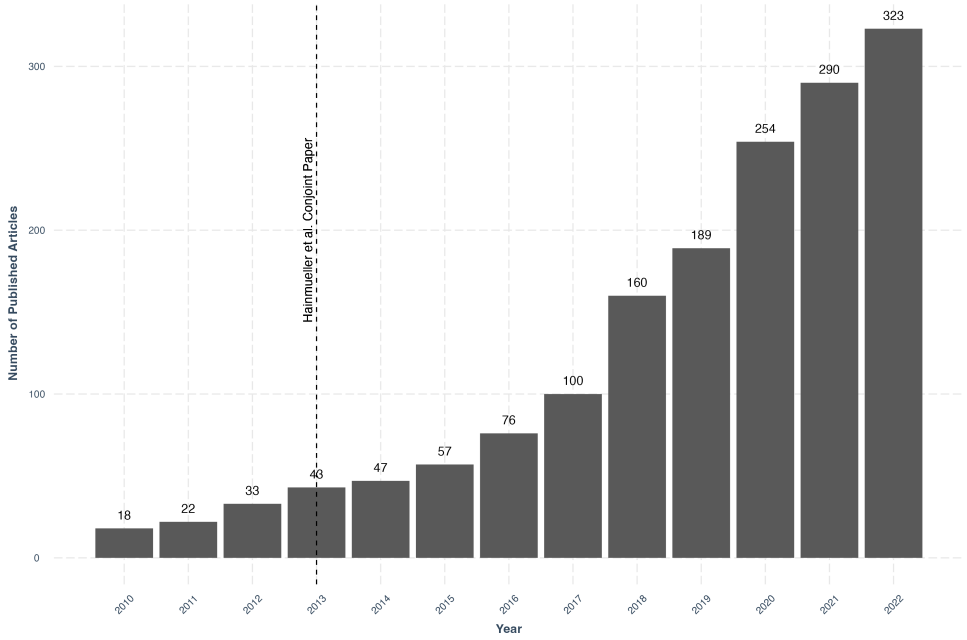

Figure 1: Recent increase of conjoint analysis in Political Science journal articles. The plot is based on data extracted from dimensions.ai, one of the largest research information databases containing information on published academic research. The articles represent the number of published articles in political science journals from 2010 to 2022 containing as a keyword "conjoint experiment" or "conjoint analysis". See also de la Cuesta, Egami, and Imai (2022) on the same topic.

The increasing use of conjoint experiments in social and behavioral sciences can be attributed to various factors. Among these, the use of computer-based surveys has had a significant impact on the development of different experimental designs (for a broader discussion on how computer-administered surveys impacted experimental research in Political Science see, Druckman et al. 2011). Due to their high dimensional nature and rather complex randomisation procedures, computer-administered surveys allow researchers to carry out conjoint experiments at lower costs compared to paper and pencil.

Conjoint experiments usually require respondents to select between two options that represent various factors. This mirrors the decision-making process that individuals face

when making political choices (e.g. when voting between two candidates). The resemblance of conjoint experiments of real-world decision-making is considered to bolster both internal and external validity when contrasted with other experimental designs. Internal validity refers to the extent to which a research study accurately establishes a cause-and-effect relationship between outcome and treatment while external validity addresses the generalizability of research findings beyond the specific context of the study. Scholars have demonstrated that conjoint designs effectively mirror real-world outcomes (e.g., Hainmueller, Hangartner, and Yamamoto 2015) and enhance realism in comparison to the direct solicitation of preferences along a single dimension (Hainmueller, Hopkins, and Yamamoto 2014).

Conjoint designs have also proven to be useful in mitigating the impact of social desirability bias, a pervasive concern in numerous subjects in the realm of social sciences. Social desirability bias entails individuals responding in ways that align with societal norms or expectations rather than providing truthful responses. The multidimensional nature of conjoint experiments, coupled with the requirement to make trade-offs among various attributes, encourages participants to prioritise their true preferences rather than focusing on attributes that may appear socially desirable. For instance, (Teele, Kalla, and Rosenbluth 2018) employed a conjoint design with the explicit aim of attenuating social desirability bias in the context of gender roles and discrimination against female political candidates. Similarly, Teele, Kalla, and Rosenbluth (2018) utilised conjoint experiments to gauge opposition to social housing within respondents' neighbourhoods, a question known to be prone to social desirability bias.

2.3 Conjoint Experiments and the Potential Outcome Framework

A characteristic of this recent body of empirical literature that relies on conjoint designs is the use of a novel statistical methodology for analysing conjoint data. In line with the broader trends in the field of political science (Druckman et al. 2011), Hainmueller,

Hopkins, and Yamamoto (2014) propose a new estimation procedure based on the potential outcomes framework of causal inference. Also known as Neyman–Rubin causal model, the potential outcomes framework is an approach to assess causality (Neyman 1923; Rubin 1974). It is based on the idea of potential outcomes under different conditions. The approach postulates that each individual or unit has two potential outcomes: the potential outcome under the treatment (denoted as Y_T) and the potential outcome under the control or non-treatment condition (denoted as Y_C). As described by Holland (1986), the fundamental challenge is that only one of these potential outcomes can actually be observed for each unit, depending on whether the unit belongs to the treatment or non-treatment group.

Within the potential outcomes framework, the identification of the causal effects can be thought of as a missing data problem that entails inferring the unseen counterfactual solely based on observed data. Under a series of assumptions[2], this problem is solved by taking the average of the outcome of the units under non-treatment and subtracting it from the average of the outcome of the treatment group. The casual effect $\widehat{\tau}$ for a simple binary treatment can be defined as

$$\widehat{\tau}_{naive} = \overline{Y_T} - \overline{Y_C} \tag{1}$$

where $\overline{Y_T}$ is the average outcome of the units in the treatment group and $\overline{Y_C}$ is the average outcome of the units in the non-treatment group.

In common experimental designs, this naive estimator represents the average causal effects of the experimental manipulation as a whole. However, in scenarios like conjoint designs—where different experimental attributes shift simultaneously—there are multiple potential outcomes embedded in each experimental task. Consequently, we need an estimator that can *decompose* this composite effect and returns the marginal effect of each attribute level on the outcome of interest. To this end, Hainmueller, Hopkins, and

[2]The key assumptions of the potential outcome framework include (a) Stable Unit Treatment Value Assumption (SUTVA), (b) Consistency, (c) Ignorability (or "no unobserved confounding"), (d) Overlap, (e) Positivity (or "strong ignorability"). Please refer to Rubin (2005) for a detailed discussion on the assumptions underlying the potential outcome framework

Yamamoto (2014) propose a new estimator called the Average Marginal Component Effect (AMCE). The AMCE represents the effect of a particular attribute level of interest against another level of the same attribute while holding equal the joint distribution of the other attributes included in the design. For simplicity, I consider a single attribute X with varying levels X_J and I keep the other experimental attributes constant. In its reduced form, the AMCE for attribute X for the level X_J is calculated as the difference in average outcome between levels X_j and X_{j-1}, averaged over all the experimental units

$$AMCE(X, X_j) = \frac{1}{n}\sum_{i=1}^{n}(Y_{ij}) - \frac{1}{n}\sum_{i=1}^{n}(Y_{ij-1}) \tag{2}$$

Here, Y_{ij} represents the outcome of respondent i for a given attribute level X_j and Y_{ij-1} represent the outcome of respondent i for another level X_{j-1} of the same attribute[3]. The formula makes clear that by marginalising the other attributes included in the design, the AMCE averages over the effect variations caused by each possible combination of the other attributes included in the design. Intuitively, the AMCE can be conceived as the mean difference in the population outcome that would result if a respondent were shown all the profiles with attributes X_j as opposed to X_{ij-1} (for a discussion on this topic see, Leeper, Hobolt, and Tilley 2019).

Hainmueller, Hopkins, and Yamamoto (2014) show that the AMCE is non-parametrically identified from conjoint data as long as (1) the levels are uniformly randomised at the profile level, (2) there is no profile- and attribute-order effects and (3) carryover effects between different experimental task are absent. Unsurprisingly, most of the postulated assumptions are similar to one needed for the identification of the Average Treatment Effect in common factorial experiments and, thus, are not discussed in the text. Here, I focus on the profile randomisation assumption since it requires a particular type of attribute randomisation compared to other conjoint designs (on this point see, Hainmueller, Hopkins, and Yamamoto 2014). This assumption postulates that the assignment of the attribute levels to the conjoint profiles seen, by each respondent, in each experimental

[3]For sake of simplicity, the equation does not take into account that the AMCE also averages over the various choice tasks that are usually included in conjoint experiments.

task, is fully randomised. This ensures that the attributes are orthogonal or independent of each other and, thus, that the outcome is uncorrelated to specific profiles presented in the experimental task. It also implies that there must be a non-zero probability of occurrence of all the possible attribute combinations for which the potential outcomes are defined[4].

Thanks to the randomisation assumption, the AMCE requires no functional form assumption and can be estimated using a linear regression model where the dependent variable is the experimental outcome (either dichotomous or continuous)[5] and the independent variables are a series of dummies representing the entire set of attributes contained in the conjoint design. Each dummy variable corresponds to a specific level of one of the conjoint attributes. The reference category for each attribute is omitted, and the model estimates coefficients for the remaining dummy variables. The proposed estimator is similar to a difference-in-means estimator where the units are collapsed into smaller groups based on the attribute level of interest (see, Leeper, Hobolt, and Tilley 2019). In this case, the coefficients can be thought of as the difference in the mean of the outcome between two of these collapsed groups.

2.4 Statistical power and Type I/II/S/M errors

In addition to the choice of the estimator, researchers need to carefully consider several factors to ensure the study's validity, reliability, and scientific relevance. One important design aspect is statistical power (Cohen 2013, for a detailed exposition of the concepts presented in this section see,). Power analysis involves assessing the probability of detecting a true effect if it actually exists. A high-powered study minimises the risk of Type II errors (*false negatives*), ensuring that the experiment can effectively detect meaningful effects. A Type II error occurs when a null hypothesis that is actually false is not rejected.

[4]In case there is dependency (i.e., non-orthogonality) between some of the attributes, the AMCE must be calculated (i.e., averaged) using appropriate weights and attributes interactions (for more details refer to, Hainmueller, Hopkins, and Yamamoto 2014)

[5]Please note that Hainmueller, Hopkins, and Yamamoto (2014) does not discuss multinomial outcomes. Please refer to Carson et al. (1994) for a discussion on multinomial outcomes in choice models.

In this case, the statistical test fails to detect a true effect. For example, if an experiment does not detect the effect of the experimental manipulation when, in fact, the treatment had an effect on the outcome, it's a Type II error.

The probability of making a Type II error is denoted as β and the statistical power of a study is usually defined as $1 - \beta$. This implies that the likelihood of making a Type II error has an inverse relationship with the power of a study: the higher the power, the lower the probability of committing a Type II error. A high probability of committing a Type II error is problematic, especially in experimental studies where researchers are interested in assessing the causal relationship between two variables. It can hide important findings, lead to accepting incorrect conclusions, and challenge findings from previous studies that are, in fact, valid (Cohen 2013). This can undermine perfectly well-grounded theories and misguide further theoretical reflections based on the (falsely) negative result obtained in the study (Breur 2016). Furthermore, studies with non-significant results are often less likely to be published in academic journals, leading to publication bias and a skewed perception of the effect size published in the literature (Baker 2016)

Another concept closely associated with power analysis and significance testing is Type I error, which occurs when a null hypothesis that is, in fact, true, is erroneously rejected. This is commonly referred to as a *false positive*: due to random variability the statistical test discovers a significant effect that doesn't actually exist. The probability of committing a Type I error is denoted as α and is typically referred to as the significance level. In this case, $1 - \alpha$ is the probability of obtaining a "true negative" or correctly not rejecting the null hypothesis. When the significance level is set at 0.05 (as commonly done in the social sciences), a Type I error is allowed to occur less than 5% of the time. Over multiple repetitions of the same experiment or study, around 5% of instances will result in a p-value below 0.05, even though the null hypothesis is true.

Similar to the relationship between power and Type II error, the connection between Type I error and power demonstrates an inverse relationship. When the probability of Type I errors decreases, the required statistical power is higher, keeping everything else

equal. Conversely, a higher likelihood of Type I errors corresponds to an increase in statistical power. This dynamic reveals the existence of trade-offs between Type I and Type II error rates. As the significance level (α) decreases in an effort to mitigate the risk of Type I errors, there is a concurrent rise in the risk of committing Type II errors, assuming other factors (such as the sample size of the study) remain constant.

Striking a balance between these two error types becomes pivotal when designing an experiment. For experiments where detecting true effects is crucial, opting for a higher significance level (α) to increase power may be beneficial. This is especially relevant when the real-world repercussions of a Type II error are substantial, such as failing to diagnose a life-threatening condition in a patient. Conversely, higher Type I error rates may be undesirable when false positives pose a significant risk, such as the approval of a new drug with severe side effects. In these instances, researchers might be willing to increase the probability of obtaining false negatives to reduce the risk of false positives.

As previously mentioned, power analysis focuses on the probability of finding an effect when it is present in the population. However, even when a study finds a statistically significant effect, a possibility that the discovered effect is biased remains present. This issue has been outlined in the work of Gelman and Carlin (2014) and further elaborated by Lu, Qiu, and Deng (2019). The authors introduce two new types of errors, namely Type S and Type M error rates. Type S errors indicate the probability that the sign of an estimated effect is incorrect. Type M errors quantify the factor by which the estimated effect might overshoot the true population parameter. Intuitively, it can be interpreted as an *exaggeration ratio* of the true coefficients[6].

If the statistical power is high enough, the incidence of Type S and Type M error is low and, thus, unproblematic for the validity of the estimated effects (Gelman and Carlin 2014). However, when an experiment does not reach adequate power (e.g., ≤ 0.80), the probability that the estimated effect sign is incorrect or the magnitude is biased increases substantially. Especially in these cases, the emphasis solely on null hypothesis

[6]For the closed-form expressions of Type S and Type M and related proofs please refer to Lu, Qiu, and Deng (2019).

significance testing can be severely limiting and, thus, these new measures can provide additional guidance for hypothesis testing and the design of an experiment (Lu, Qiu, and Deng 2019). Type S errors are especially relevant when the direction of the estimated effect is of prime importance. This can be the case when researchers are interested in testing novel cause-and-effect mechanisms where the direction of hypothesised effect is not clear. Type M errors become relevant when the size of the effect size is of prime importance such as in fields where even small changes can have considerable real-world consequences.

3 Power Analysis in Conjoint Experiments

In traditional experiments, statistical power (and related measure proposed by (Gelman and Carlin 2014)) is dependent on the desired trade-off between Type I and Type II errors, the sample size, and the effect size. Conjoint experiments, similar to other factorial designs, add another layer of complexity in terms of design choices. In addition to sample and effect size considerations, experimentalists need to decide on the number of attributes (the features of the profiles, such as the gender of a hypothetical candidate), the number of levels of such attributes (the values of that attributes can take, such as male or female), and the number of tasks (the number of profile choices a respondent will undertake).

Previous research on sample size requirements for conjoint experiments is slim and fails to provide accurate guidance for designing conjoint experiments (Orme 1998; Jordan J. Louviere et al. 2000; Rose and Bliemer 2013). The few works on the topic do not directly focus on the fully randomised approach proposed by (Hainmueller, Hopkins, and Yamamoto 2014) and typically disregard the number of attributes and levels included in the design (e.g., Jordan J. Louviere et al. 2000; Rose and Bliemer 2013). In addition, current literature provides no indications of the impact of design choices on the probability of an estimate being biased upward or downward (Type M error) or being in the wrong direction (Type S error).

The sole exception is the working paper authored by Schuessler and Freitag (2020),

which emerged subsequent to the working paper by Stefanelli and Lukac (2020) which draws upon the initial idea presented in this thesis (refer to the *Contribution Statement* section for more details). Schuessler and Freitag (2020) employ non-parametric techniques to derive minimal sample size requirements for the conjoint design proposed by Hainmueller, Hopkins, and Yamamoto (2014). Although Schuessler and Freitag (2020)'s paper offers valuable guidance for simple conjoint designs, it is limited in two notable aspects. Namely, the paper omits the cluster structure typically found in conjoint designs and disregards the potential trade-off between the number of tasks and the number of respondents. In contrast with what has been suggested by previous literature (van Breukelen and Candel 2012; Moerbeek 2011; M. Chang and Chow 2006), this choice implies that having 500 respondents complete 10 tasks equates 5000 respondents completing only 1 task[7] Second, their approach cannot easily be expanded to accommodate more complex designs or sampling strategies such as multi-stage sampling, or when individual-level (or cluster-level) heterogeneity is expected to be more pronounced. This limitation extends beyond Schuessler and Freitag (2020)'s work and encompasses all non-parametric techniques for power analysis in contexts where no straightforward mathematical relationships exist between various aspects of the design.

Therefore, this thesis develops a flexible simulation-based framework based on the approach proposed by Hainmueller, Hopkins, and Yamamoto (2014). In addition to the derivation of the minimal required sample size given the number of respondents and experimental conditions, this work adds to the existing literature by providing guidance on the exchangeability between respondents and repeated tasks. In applied research, this can be used to obtain useful insights on whether designs with a low number of respondents who fulfil a large number of tasks are as good as designs with a larger number of respondents who performs a lower number of tasks. Given a range of pre-specified effect sizes, the proposed simulation also allows researchers to easily adjust the degree of confidence of the estimated parameters and, thus, directly evaluate the trade-off between the false-positive

[7]In the context of conjoint experiments, this is especially relevant when the number of tasks per individual is large and the number of attributes and attributes levels included in the design is relatively small Abadie et al. (2023). See the Conclusions for more details.

and false-negative. Lastly, given the flexible nature of our data generating process, our simulation can be easily modified to accommodate more complex theoretical scenarios or sampling strategies (see Conclusions for more information on this point). For instance, similarly to what has been proposed in Schuessler and Freitag (2020), the R code made public with this thesis can be used to calculate statistical power when there is treatment effect heterogeneity across different observed subgroups of respondents.

3.1 Data-generating process

In order to estimate the power of a wide range of conjoint experimental designs, I formulate an underlying data-generating model that addresses the shortcomings identified in the previous section while retaining the necessary versatility to accommodate a vast array of designs employed in empirical research. I start with a basic conceptual model of conjoint decision-making and sequentially expand the model to its final shape. For the sake of simplicity, I focus on the most common design employed in the literature, a choice-based conjoint experiment with two profiles.

Starting with a hypothetical example of a conjoint in Figure 2, any choice-based conjoint experiment consists of profiles, e.g. two candidates that a respondent can choose from (selection of a profile is denoted as $Y_{p_1} = 1$, while $Y_{p_1} = 0$ when a profile is not selected). These profiles systematically differ in a set of attributes (denoted as $X = \{X_1, ..., X_k\}$); in this example, gender, level of education, and political leaning are considered attributes. Each of the k attributes has a number of l levels. For example, the attribute representing the candidate's gender (X_1) contains two levels: female and male ($X_1 \in \{Female, Male\}$).

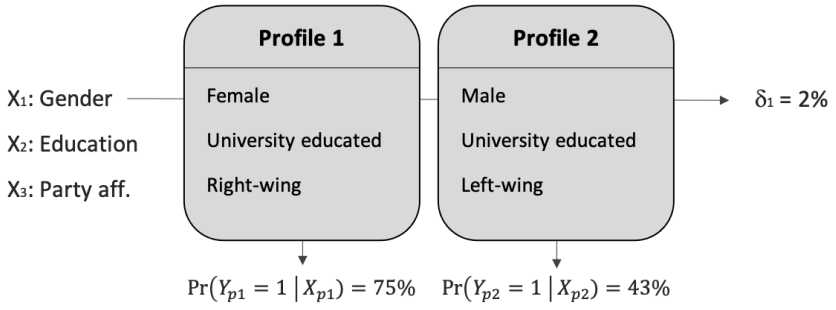

Figure 2: A hypothetical conjoint experimental task

The model proposed by Hainmueller, Hopkins, and Yamamoto (2014) is agnostic to the type of decision-making process that respondents undertake when choosing between two profiles. Their approach is based on the probability of selecting a profile given its attributes and ignores how these attributes are weighted against each other, or whether respondents perform any type of trade-offs between the different profiles. However, for the purposes of this study, a data-generating model is needed for calculating the statistical power of various conjoint designs. This requires establishing a minimal set of assumptions regarding the behavioural model that respondents use when choosing between two profiles.

Based on previous research (Jenke et al. 2020; Meißner and Decker 2010), it is assumed that (most) respondents form the probability of selecting a profile in two complementary ways. Respondents make *within-attribute* comparisons—i.e. attribute by attribute. Using the example above, a respondent slightly preferring female candidates over males, would increase the probability of selecting Profile 1 over Profile 2 after seeing that the former contains a female, while the latter is a male candidate. Then, they move to the evaluation of the second attribute, and so forth. At the same time, respondents form an underlying latent probability of selecting a profile, which I call *within-profile* decision-making. A respondent would evaluate profiles sequentially and individually, forming a probability of selecting each profile given its attributes. In this stylised example, given the attributes in Profile 1 (X_{p_1}), the respondent would choose this profile with a probability of 75%, and given the attributes in Profile 2 (X_{p_2}), the respondent would choose this profile with a probability of 43

The probability that a profile would be selected over another randomly chosen profile is therefore defined as $Pr(Y_{p_1} = 1|X_{p_1})$, where X_{p_1} are the attributes of the profile: e.g. Profile 1 has attributes X_1 = "Female", X_2 = "University educated", X_3 = "Right-wing". Importantly, the within-profile selection probability is expected to be consistent with the within-attribute causal effects. However, the quantity of interest in conjoint experiments is not the within-/between- attribute effects but, rather, the AMCE of each attribute included in the design, while controlling for all other attributes. In terms of

the data-generating process, it is not possible to generate the AMCEs directly since these latent underlying probabilities do not sum to 1. This prevents an immediate closed-form expectation about profile selection. One of the approaches, identified by Gall (2020), proposes the following: start with a 50% probability of selecting either of the profiles and adjust this probability by the causal effect divided by two (for two profile alternatives): e.g. the probability of selecting Profile 1 given its gender (assuming $\delta_1 = 2\%$) would increase to $51\% = 50\% + 2\%/2$. This adjustment guarantees that the resulting probabilities sum to 1.

Despite the validity of the approach proposed by Gall (2020), the aim of this thesis is to produce a model that is flexible to any number of attributes or levels and does not require adjusting the probability of selection based on the various δ. To this end, I combine selection probabilities into one number to run a Bernoulli trial on the selection in a hypothetical conjoint experiment. The interest lies not in $Pr(Y_{p_1} = 1|X_{p_1})$, i.e. the probability of choosing a profile given its attributes, but rather in $Pr(Y_{p_1} = 1|X_{p_1}, X_{p_2})$, the probability of choosing Profile 1, given the attributes of Profile 1 and Profile 2. As can be seen, $Pr(Y_{p_1} = 1|X_{p_1}, X_{p_2}) = 1 - Pr(Y_{p_2} = 1|X_{p_1}, X_{p_2})$, hence, fulfills the condition of summing to 1. It is solved by transforming individual probabilities to odds ratios in the following equation:

$$OR(\text{P1 vs. P2 } |X_{p1}, X_{p2}) = \frac{\text{odds } P_1}{\text{odds } P_2} = \frac{\frac{Pr(Y_{p_1} = 1|X_{p_1})}{Pr(Y_{p_1} = 0|X_{p_1})}}{\frac{Pr(Y_{p_2} = 1|X_{p_2})}{Pr(Y_{p_2} = 0|X_{p_2})}} \tag{3}$$

$$OR(\text{P1 vs. P2 } |X_{p1}, X_{p2}) = \frac{\text{odds } P_1}{\text{odds } P_2} = \frac{\frac{Pr(Y_{p_1} = 1|X_{p_1})}{Pr(Y_{p_1} = 0|X_{p_1})}}{\frac{Pr(Y_{p_2} = 1|X_{p_2})}{Pr(Y_{p_2} = 0|X_{p_2})}} \tag{4}$$

In a subsequent step, $OR(\text{P1 vs. P2}|X_{p1}, X_{p2})$—the odds of selecting Profile 1 over selecting Profile 2, given attributes of both profiles—is transformed back to probability using the following equation:

$$Pr(Y_{p_1} = 1|X_{p1}, X_{p2}) = \frac{OR(\text{P1 vs. P2 } |X_{p1}, X_{p2})}{1 + OR(\text{P1 vs. P2 } |X_{p1}, X_{p2})} \tag{5}$$

The resulting probability $Pr(Y_{p_1} = 1|X_{p1}, X_{p2})$ is the probability of selecting Profile 1, given the attributes of Profile 1 and Profile 2. Finally, I can generate the outcome variable of a hypothetical conjoint experiment as a random draw from a Bernoulli distribution with $p = Pr(Y_{p_1} = 1|X_{p1}, X_{p2})$.

3.2 Model for conjoint experiments

In order to simulate the data for power analysis, we need to design a statistical model that yields the main input into the previous section: $Pr(Y_{p_1} = 1|X_{p1})$, namely the probability of selecting a profile given its attributes. I assume that the probability of choosing a profile is a linear combination of levels (i.e. treatments) shown in the profile (on this point see also, Jordan. J. Louviere 2015). I use a linear probability model to define $Pr(Y_{p_1} = 1|X_{p1})$. Using a logistic function to generate the data is a viable alternative and the simulation model can be extended in that direction. However, the linear probability model with imposed minimum and maximum is a good approximation of the logistic function. Furthermore, based on the results of the literature review, most of the discovered AMCEs are between 0.05 and 0.16 making the selection probability well within the [0, 1] range. Lastly, using a linear probability model correspond to the intuitive interpretation of the AMCE proposed by Hainmueller, Hopkins, and Yamamoto (2014), which can facilitate the understanding of our approach in the applied research community. The model is formulated as

$$Pr(Y_{p_1} = 1|X_{p1}) = \beta_0 + \sum_{k} \sum_{(l-1)} \beta_{kl} X_{kl} \tag{6}$$

where β_0 is the intercept—the probability of selecting a profile that has attributes corresponding to the reference categories of all attributes, and β_{kl} is the treatment effect of level l of variable k, relative to the reference category (one category is always omitted as a reference, denoted by $l - 1$ under the summation).

As mentioned before, I have to take an additional step and constrain the generated selection probabilities from the linear probability model to make sure we get valid probabilities bounded between 0 and 1. Although the limit is commonly placed at [0, 1], to overcome problems in Eq. 1 due to division by 0, I use an arbitrary clip at [0.001, 0.999]:

$$Pr(Y_{p_1} = 1|X_{p1}) = \begin{cases} 0.999 & \beta_0 + \sum_k \sum_{(l-1)} \beta_{kl} X_{kl} > 0.999 \\ \beta_0 + \sum_k \sum_{(l-1)} \beta_{kl} X_{kl} & \beta_0 + \sum_k \sum_{(l-1)} \beta_{kl} X_{kl} \in [0.001, 0.999] \\ 0.001 & \beta_0 + \sum_k \sum_{(l-1)} \beta_{kl} X_{kl} < 0.001 \end{cases} \tag{7}$$

In the final step, I extend the structural model (Eq. 3 and Eq. 4) by including random effects (Train 2009). This allows for the possibility of treatment heterogeneity across respondents. By treatment heterogeneity, I mean that each respondent can have underlying characteristics that impact the probability of choosing one profile given its attributes and levels. Such respondent-specific characteristics can be observed (e.g. ethnicity, gender, age or political preference of the respondent) or unobserved (i.e. questions non-included in the survey) and can take various functional forms. For the purposes of this analysis, I use a normally distributed random effect for each of the β_{kl} nested within respondents indexed by j (see Conclusion for a discussion on this point):

$$\begin{aligned} Pr(Y_{p_1} = 1|X_{p1}) &= \beta_0 + \sum_k \sum_{(l-1)} \beta_{klj} X_{kl} \\ \beta_{klj} &= \gamma_{kl0} + u_{klj} \\ u_{klj} &= N(0, \Sigma) \end{aligned} \tag{8}$$

where γ_{kl0} is a slope constant with respondent-level deviations u_{klj} and Σ is a variance-covariance matrix with $\sigma_1^2 ... \sigma_{k(l-1)}^2$ on the diagonal and zeroes elsewhere. With this formulation, I explicitly add nesting of trials within individuals, where each attribute and level has its respondent-specific error term. This provides a simple model that allows heterogeneity across treatments and potential expansions to simulate the presence of unobserved subgroups with significantly different treatment effects. This assumes that the random effects are mutually uncorrelated, however, it can be expanded by adding covariances of

error terms and sampling from a multivariate normal distribution instead.

4 Setup of the simulation and evaluation criteria

4.1 Literature review of conjoint studies in political science

As previously mentioned, conjoint experiments add several layers of complexity to the calculation of power due to the higher number of features: the sample size, the number of tasks performed by each respondent, the number of attributes and levels of each attribute, and the size of the hypothesised effect in the population. To ensure that the parameters utilised for power simulation align with the design choices seen in conjoint designs, it is required to understand how conjoint experiments are conducted in political science. This means that the simulation results can be used to assess the power of common conjoint designs and that the simulation remains computationally manageable. To this end, I carried out an extensive literature review focusing on each of these design considerations. The data collection resulted in a total of 59 articles from 2014 to 2020 (see Table 2 for the complete list of papers). Although the list is not exhaustive of the entire conjoint literature, it provides a sufficiently detailed understanding of how conjoint experiments are designed and implemented in political science and neighbouring disciplines.

The data collection for the literature review consisted of two different phases. First, I selected 14 journals in the field of political science (Table 1) and I performed a keyword search using the term "conjoint". Second, I added articles that have cited Hainmueller, Hopkins, and Yamamoto (2014) work using the "cited by" feature in Google Scholar. I removed all articles that were not using primary sources or whose contributions were predominantly methodological. I also removed those articles that were clearly outside of the field of political science and neighbouring disciples. To make the results comparable, I restricted the literature review to forced-choice conjoint experiments that use the AMCE as defined by Hainmueller, Hopkins, and Yamamoto (2014). Articles using rating scales (e.g., Huff and Joshua D. Kertzer 2018) or logit-based estimators (Duch et al. 2020, e.g.,

) have been also excluded.

Journal Name
American Political Science Review
American Journal of Political Science
The Journal of Politics
British Journal of Political Science
European Journal of Political Research
Political Research Quarterly
Perspectives on Politics
Political Behavior
Public Opinion Quarterly
European Political Science Review
Electoral Studies
Journal of Experimental Political Science
Political Science Research and Methods
Research & Politics

Table 1: List of included journals

I coded each article along several dimensions, namely the number of respondents, the total number of conjoint tasks that the respondents were asked to complete, the total number of observations, the number of attributes, the number of levels of the attribute with the most categories (i.e., levels), the AMCEs, and whether their respective p-values were significant at $\leq .05$. First, I collected information on the sample size. Specifically, I coded the number of respondents, the total number of conjoint tasks that the respondents were asked to complete, and the total number of observations (effective sample size) employed in the analysis. In case the number of tasks was not directly reported in the article, I first consulted the article's supplementary materials. If they did not contain sufficient information to determine the number of tasks, I estimated the number of tasks by dividing the reported number of observations by the total number of respondents. In case multiple samples were collected to validate the experimental results, only the sample used in the main text of the article has been coded (e.g., Bechtel and Liesch 2020). On the contrary, if an article included different conjoint experiments, I generated an entry for each study [e.g., (e.g., Peterson and Simonovits 2018).

Title	Author(s)	Year
Conditional legitimacy: How turnout, majority size, and outcome affect perceptions of legitimacy in European Union membership referendums	Arnesen, Broderstad, Johannesson, & Linde	2019
Fit for the Job: Candidate Qualifications and Vote Choice in Low Information Elections	Atkeson & Hamel	2020
How Clients Select Brokers: Competition and Choice in India's Slums	Auerbach & Thachil	2018
The Structure of American Income Tax Policy Preferences	Ballard-Rosa, Martin, & Scheve	2017
Interests, Norms and Support for the Provision of Global Public Goods: The Case of Climate Co-operation	Bechtel, Genovese, & Scheve	2019
Policy Design and Domestic Support for International Bailouts	Bechtel, Hainmueller, & Margalit	2017
Reforms and Redistribution: Disentangling the Egoistic and Sociotropic Origins of Voter Preferences	Bechtel & Liesch	2020
Attribute Affinity: U.S. Natives' Attitudes Toward Immigrants	Berinsky, Rizzo, Rosenzweig, & Heaps	2020
Legislator Dissent as a Valence Signal	Campbell, Cowley, Vivyan, & Wagner	2019
Why Friends and Neighbors? Explaining the Electoral Appeal of Local Roots	Campbell, Cowley, Vivyan, & Wagner	2019
Who wants to hire a more diverse faculty? A conjoint analysis of faculty and student preferences for gender and racial/ethnic diversity	Carey et al.	2018
Do Voters Dislike Working-Class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class	Carnes & Lupu	2016
Getting Rich Too Fast? Voters Reactions to Politicians Wealth Accumulation	Chauchard, Klašnja, & Harish	2019
Reciprocity and Public Opposition to Foreign Direct Investment	Chilton, Milner, & Tingley	2017
How citizens evaluate participatory processes: a conjoint analysis	Christensen	2020
Exposure to Immigration and Admission Preferences: Evidence from France	Clayton, Ferwerda, & Horiuchi	2019
Confident and cautious candidates: Explaining under-representation of women in Danish municipal politics	Dahl & Nyrup	2020
Do Local Party Chairs Think Women and Minority Candidates Can Win? Evidence from a Conjoint Experiment	Doherty, Dowling, & Miller	2019
Corruption, Accountability, and Gender: Do Female Politicians Face Higher Standards in Public Life?	Eggers, Vivyan, & Wagner	2018
Multi-dimensional preferences for labour market reforms: a conjoint experiment	Gallego & Marx	2017
What Goes with Red and Blue? Mapping Partisan and Ideological Associations in the Minds of Voters	Goggin, Henderson, & Theodoridis	2019
Institutional reform and public attitudes toward EU decision making	Hahm, Hilpert, & König	2020
The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants	Hainmueller & Hopkins	2015
Multiple Dimensions of Bureaucratic Discrimination: Evidence from German Welfare Offices.	Hemker & Rink	2017
Policy Preferences and Policy Legitimacy After Referendums: Evidence from the Brexit Negotiations	Hobolt, Tilley, & Leeper	2020
Identifying voter preferences for politicians,Äô personal attributes: a conjoint experiment in Japan	Horiuchi, Smith, & Yamamoto	2018
Measuring Voters' Multidimensional Policy Preferences with Conjoint Analysis: Application to Japan' 2014 Election	Horiuchi, Smith, & Yamamoto	2020
Citizens' Attitudes to Contact Tracing Apps	Horvath, Banducci, & James	2020
How the Public Defines Terrorism	Huff & Kertzer	2018
Moderation and Competence: How a Party's Ideological Position Shapes Its Valence Reputation	Johns & Kölln	2019
How Do Observers Assess Resolve?	Kertzer, Renshon, & Yarhi-Milo	2019
Candidate Choice Without Party Labels	Kirkland & Coppock	2018
When Do Voters Sanction Corrupt Politicians?	Klašnja, Lupu, & Tucker	2020
More Important, but for What Exactly? The Insignificant Role of Subjective Issue Importance in Vote Decisions	Leeper & Robison	2020
Do Voters Prefer Just Any Descriptive Representative? The Case of Multiracial Candidates	Lemi	2020
The Logic of Authoritarian Political Selection: Evidence from a Conjoint Experiment in China	Liu	2019
Voting for the lesser evil: evidence from a conjoint experiment in Romania	Mares & Visconti	2020
Biases at the Ballot Box: How Multiple Forms of Voter Discrimination Impede the Descriptive and Substantive Representation of Ethnic Minority Groups	L. Martin	2019
All Sins are not Created Equal: The Factors that Drive Perceptions of Corruption Severity	N. Martin & Blinder	2020
Multi-dimensional policy preferences in the 2015 British general election: A conjoint analysis	Matsuo & Lee	2018
Citizen preferences about border arrangements in divided societies: Evidence from a conjoint experiment in Northern Ireland	Morgan-Jones, Sudulich, Cochrane, & Loizides	2020
News from the Other Side: How Topic Relevance Limits the Prevalence of Partisan Selective Exposure	Mummolo	2016
Why Partisans Do Not Sort: The Constraints on Political Segregation	Mummolo & Nall	2017
Merit, Tenure, and Bureaucratic Behavior: Evidence From a Conjoint Experiment in the Dominican Republic	Oliveros & Schuster	2018
Do voters prefer gender stereotypic candidates? evidence from a conjoint survey experiment in Japan	Ono & Burden	2019
The Contingent Effects of Candidate Sex on Voter Choice	Ono & Yamada	2020
The Role of the Information Environment in Partisan Voting	Peterson	2017
The Electoral Consequences of Issue Frames	Peterson & Simonovits	2018
The Organizational Voter: Support for New Parties in Young Democracies	Poertner	2020
How Fair Is It? An Experimental Study of Perceived Fairness of Distributive Policies	Rodon & Sanjaume-Calvet	2020
Ambitious Women: Gender and Voter erceptions of Candidate Ambition	Saha & Weeks	2020
Ideology and Vote Choice in U.S. Mayoral Elections: Evidence from Facebook Surveys	Sances	2018
From ,"Different" to "Similar": An Experimental Approach to Understanding Assimilation	Schachter	2016
(Sex) Crime and Punishment in the #MeToo Era: How the Public Views Rape	Schwarz, Baum, & Cohen	2020
Political Considerations in Nonpolitical Decisions: A Conjoint Analysis of Roommate Choice	Shafranek	2019
The public view of immigrant integration: multidimensional and consensual. Evidence from survey experiments in the UK and the Netherlands	Sobolewska, Galandini, & Lessard-Phillips	2017
The importance of input and output legitimacy in democratic governance: Evidence from a population-based survey experiment in four West European countries.	Strebel, Kübler, & Marcinkowski	2019
The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics	Teele, Kalla, & Rosenbluth	2018
House or home? Constituent preferences over legislator effort allocation: Constituent Preferences Over Legislator Effort Allocation	Vivyan & Wagner	2016
Public Attitudes toward Young Immigrant Men	Ward	2019

Table 2: List of articles included in the literature review

Second, I counted the number of attributes, the number of levels of the attribute with the most categories, and the total number of levels that were part of the experimental design. This information is used to assess how large is the experimental design $J \times K$ where J is the number of attributes, K is the number of levels of the attribute with the highest number of categories. Lastly, I coded the AMCEs and their respective significance level. I excluded any Average Component Interaction Effects (ACIE) for the attributes specified in the analysis[8]. I first aimed to extract the estimates from regression tables usually contained in the supplementary materials. If no regression tables were available, I visually inspected the dot-and-whisker plots included in the main text to approximate the AMCE effect size[9]. Lastly, I coded whether the reported AMCEs were significant using either the regression table or, when not available, whether the confident interval included the zero line in the dot-and-whisker plots[10].

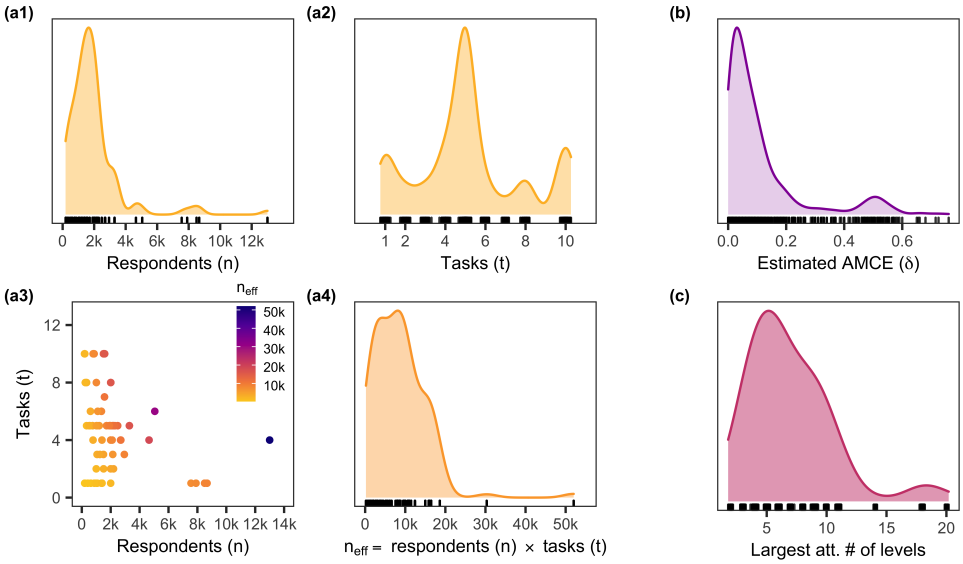

Figure 3: Literature review analysis

The literature review yielded the results in Figure 3. The median sample size in

[8]The ACIE is of interest when analysing treatment effect heterogeneity across different subgroups of respondents. Please refer to the Conclusions for more info on interaction effects and power analysis.

[9]Deriving the estimated coefficients from dot-and-whisker plots is prone to coding errors. A Research Assistant has been tasked to independently code all the AMCEs in those articles that only reported dot-and-whisker plots. Only 5 AMCEs were miscoded and were subsequently corrected.

[10]I initially planned to collect data on t and p values associated with the AMCEs. However, in many of the articles, no regression tables were available and I had to resort to the visual inspection of dot-and-whisker plots to obtain the AMCE significance. Consequently, collecting t and p values for many of the included articles was impossible.

the reviewed articles is equal to 1583 respondents (25^{th} percentile: $p_{25} = 1020$ and 75^{th} percentile: $p_{75} = 2144$) with approximately 5 tasks per respondents ($p_{25} = 4$ and $p_{75} = 6$). Panel (a3) in Figure 3 shows that the number of respondents and tasks are used as substitutes, and studies with a low number of respondents tend to compensate with a higher number of tasks given to each respondent. The sample size (n_{eff} = number of respondents (n) $\times$ number of tasks (t)) used for the analysis ranged widely between 200 and 52000 ($p_{25} = 3171$ and $p_{75} = 11000$). The median number of attributes is 7 ($p_{25} = 6$ and $p_{75} = 8$) and seems to follow what is recommended by recent literature on satisficing in conjoint experiments (Bansak et al. 2019). The median of the number of levels for the biggest attribute is 6 ($p_{25} = 4$ and $p_{75} = 9$). The median AMCE is equal to 0.05 ($p_{25} = 0.025$ and $p_{75} = 0.12$). The 25^{th} percentile of the *significant* AMCEs is and 0.049, the median is 0.09, and the 75^{th} percentile is 0.16. This means that about 25% of the significant AMCEs lie between 0.005 (the minimum) and 0.049 and 50% of the effects are below 0.09. The median of the *insignificant* AMCEs is 0.02 ($p_{25} = 0.01$ and $p_{75} = 0.03$). These results indicate that a large portion of the effects found in conjoint experiments are relatively small. Only 40% of the reviewed studies employed a probability-based representative sample while the rest relied on a convenience sample (usually stratified by gender, education, and other relevant census variables).

4.2 Simulation set up

The simulation is set up with the experimental conditions shown in Table 3 which were derived from the literature review presented in the previous section (see Figure 3). It was performed using the entire range of combinations for each element included in Table 3. This resulted in a total of 2520 conditions ($8 \times 5 \times 7 \times 9$) which were replicated 1000 times each.

As indicated in Table 3, the number of attributes is kept constant while the number of levels varies across the different experiment conditions. This is due to the fact that as the number of attribute levels increases, the number of respondents used for the calculation

of the AMCE decreases. For instance, if an experiment includes a single attribute with only two levels, the AMCE is estimated using $N\frac{1}{2}$ observations. On there other hand, if an experiment includes an attribute with 8 levels, the AMCE is calculated using $N\frac{1}{8}$ observations. This results in a loss of precision of estimated effects and, consequently, lower statistical power.

This is because, by design, the AMCE (δ) represents the causal effect of a single profile attribute while marginalising all the other attributes included in the design. The marginalisation is a fundamental feature of the design proposed by Hainmueller, Hopkins, and Yamamoto (2014) and, as mentioned before, it is made possible by the fact that the attributes are randomised independently at the profile level. Thanks to the marginalisation, the addition of one or more attributes to the design does not change the number of respondents that are *treated* with each level. Consequently, the number of attributes has no substantial impact on power calculations for conjoint experiments with orthogonal attributes and has been kept constant in the simulation (Hainmueller, Hopkins, and Yamamoto 2014).

Based on the recommendation in the literature, I also use clustered standard errors to calculate all three quantities (Arceneaux 2005; Hainmueller, Hopkins, and Yamamoto 2014). Nevertheless, since the simulation procedure fixes the treatment heterogeneity between respondents $\sigma_k = 0.02$, the simulation is unable to test scenarios where respondent-level deviations u_{klj} are more pronounced. I also fixed the number of profiles to 2 as it is common for conjoint experiments that follow Hainmueller, Hopkins, and Yamamoto (2014).

It is worth noting that in forced conjoint experiments the AMCE (δ) represents the probability change of choosing one profile given its attributes and not the standard deviation of the outcome from the sample average (standardised effect size), such as the commonly used *Cohen's d*. In the text, I do not focus on Cohen's d since it is not usually employed in conjoint experiments that employ the Hainmueller, Hopkins, and Yamamoto (2014) framework. However, there is a functional relationship between Cohen's d and

AMCE. In paired conjoint design with 2 profiles, the latent probability of choosing a profile is always, by design, 50%. Consequently, an AMCE of .1 corresponds to Cohen's d of $\frac{.1}{.5} = .2$.

Parameter	**Values**
Variable	
Respondents	500, 750, 1000, 1250, 1500, 1750, 2000, 3000
Tasks	1, 3, 5, 7, 9
Levels	2, 3, 4, 5, 10, 15, 20
Effect size δ (AMCE)	0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15, 0.20, 0.25
Constant	
Profiles	2
Attributes	5
Treatment heterogeneity σ_k	0.02
Significance level α	0.05
Satisfactory Power $(1-\beta)$	0.80

Table 3: Simulation inputs

4.3 Evaluation criteria

Regarding the evaluation criteria of the simulation, statistical power is defined as the probability to detect a non-zero population effect for a binary hypothesis test where $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$. This is equal to the probability that a statistical test will correctly reject the null hypothesis (H_0) when it is false. I assume that the test statistic follows a normal distribution $Z|\mu, \sigma \sim N(\mu, \sigma^2)$ given $\sigma > 0$. The key idea is that the significance level α is given as $Pr(|Z| > z_\alpha|H_0)$, hence the result is considered significant (H_0 is rejected) if the estimated effect surpasses some two-sided critical threshold value. Under a non-zero μ, the power function is $p = Pr(\text{reject } H_0|H_1 \text{ is true})$ or, specifically, $Pr(|Z| > \sigma z_\alpha|\mu)$. In this work, I adopt an *a priori* perspective: the goal is to achieve a pre-specified desirable power level, given a pre-specified level of α. In line with what is commonly accepted by the research community (e.g., Cohen 2013), I consider a power of 0.80 satisfactory with an α of 0.05 and β of 0.20.

As mentioned before, the use of power analysis with emphasis solely on null hypothesis significance testing can be limiting. Consequently, following Gelman and Carlin (2014),

I also calculate the probability of an estimate being in the wrong direction (the Type S error) and the factor by which the magnitude of an effect might be overestimated (the Type M error). For a non-zero μ, I define:

- *Type S error*: Probability of a Type S error is given by $s = Pr(Z < 0|\mu, |Z| > \sigma z_\alpha)$. In words, this is a probability that an estimated significant effect will carry an opposite sign than the population parameter.
- *Type M error*: Expected Type M error is given by $m = \mathrm{E}[|\hat{\mu}| \;\big|\; \mu, |Z| > \sigma z_\alpha]/\mu$.

5 Results

5.1 Sample size: respondents and tasks

In calculating power for conjoint experiments, similar to other multi-trial or clustered designs, it is necessary to choose not only the number of respondents but also the number of trials (i.e., tasks) that each respondent performs. Previous studies on cluster randomised trials (e.g., van Breukelen and Candel 2012) (number of clusters versus cluster size), longitudinal intervention studies (e.g., Moerbeek 2011) (number of subjects versus the number of measures per subject), and toxicity studies (e.g., M. Chang and Chow 2006) (number of dose levels versus the number of subjects per dose level) suggest that the number of subjects has a larger impact on power considerations. However, often the number of trials is used as a substitute for respondents in conjoint studies (see Literature Review). As no simple mathematical relationship for this trade-off exists, it is not clear how they affect the expected power of a conjoint experiment.

Figure 4 indicates that the statistical power of the AMCE is a function of (a) the number of respondents, (b) the number of trials completed by each respondent, and (c) the true effect size. I first focus on the total sample size used for the analysis, which is obtained by multiplying the number of trials by the number of respondents ($n_{eff} = n \times t$). To ease the interpretation, we hold the number of levels constant at $l = 2$. For an effect size

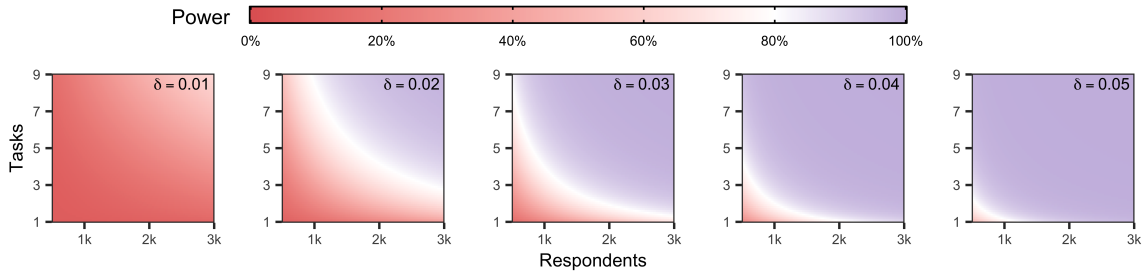

Figure 4: Effect of sample size—the number of respondents and tasks—on statistical power for different assumed AMCEs (δ). Shown values reflect a design with 2 levels ($l = 2$).

of $\delta = 0.01$—meaning shifting a profile's attribute from A to B increases the propensity of choosing a profile by 1 percentage point—the conventional power threshold of 0.80 is not reached in any of the simulated conditions. For an effect of $\delta = 0.03$, the simulation shows that some combinations of sample size (n) and the number of trials (t) of approximately $n_{eff} = n \times t = 10,000$ observations are required to achieve a well-powered design. To achieve desirable power levels for a hypothesised effect size of $\delta = 0.05$, relatively small samples are required: a sample of approximately 1,000 individuals who complete 2 tasks is sufficient to provide a well-powered design $n_{eff} = n \times t = 2000$. Lastly, with large effect sizes of $\delta \geq 0.10$, the simulation shows that conjoint experiments reach the conventional power threshold with small samples ($n = 300$) and a low number of tasks (e.g. $t = 2$).

Next, the trade-off between the number of respondents (n) and trials (t) was analysed. The results revealed that the trade-off between respondents (n) and the number of trials (t) follows an exponential distribution. For an average effect of $\delta = 0.05$ in experiments with $n = 500$ that aim to keep the assumed statistical power above 0.80, each additional task corresponds to approximately 800 respondents. For designs with $n = 1000$, the simulation shows that each additional task corresponds to approximately 2200 respondents. This means that the impact of an additional conjoint task marginally diminishes for each additional respondent in the sample. This finding is in line with previous literature on randomised trials and confirms that for designs with a lower number of respondents, more tasks are required to keep the assumed statistical power above 0.80. In other words, having 500 respondents complete 10 tasks is not as good as 5000 respondents completing

only 1 task. These indications can be used by researchers to decide on the number of tasks to include in an experiment given a fixed number of respondents.

5.2 Number of levels

Another point of interest in calculating the statistical power of a conjoint experiment is the required sample size given the number of levels of the biggest attribute included in the design. In other words, it refers to how many levels can be included in the design based on the number of respondents and the number of trials. Research on factorial designs has shown that the higher the number of levels, the lower the power, keeping the sample size constant (e.g., Dziak, Nahum-Shani, and Collins 2012). However, contrary to more traditional experimental designs, conjoint studies typically include a disproportionately large number of levels per attribute that may lead to severely underpowered designs. To gauge the impact of including different numbers of levels in a conjoint experiment, the AMCE is fixed to $\delta = 0.05$ such that it is possible to observe the trade-off between the number of levels and the pre-defined power threshold (ge 0.80).

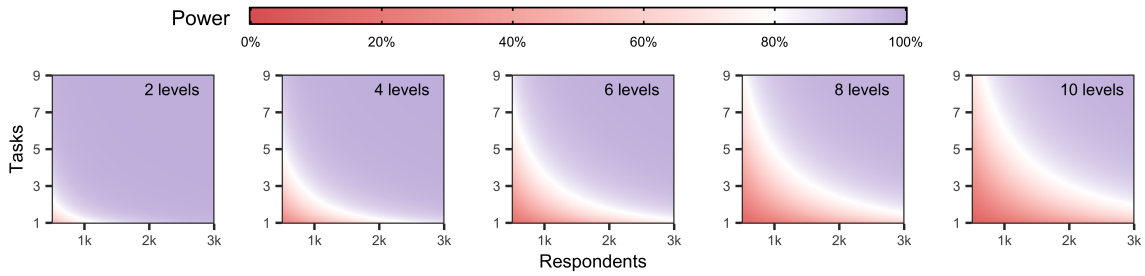

Figure 5: Effect of number of levels (l) of the largest attribute on statistical power. Shown values assume a true AMCE (δ) = 0.05.

In line with previous considerations, Figure 5 reveals that the required sample size increases as the number of levels included in a conjoint experiment increases. A design that includes at most a dichotomous attribute ($l = 2$) reaches the required power of .80 with 3 tasks distributed across 1000 respondents (assuming $\delta = 0.05$). On the contrary, using an attribute with 5 or more levels ($l \geq 5$), the same number of respondents and trials would yield a clearly underpowered experimental design. This is not surprising. As

mentioned before, this pattern arises due to the increase of the number of experimental conditions for each additional level included in the design and, consequently, a lower number of respondents being *treated* with each level.

By varying the number of levels, the simulation reveals another layer of complexity to optimise for when aiming at obtaining reliable estimates of population parameters with conjoint experiments. Although collapsing categories into a smaller number of levels may be a quick solution, it might not be an option for some research questions that aim at assessing the effects of many different categories. For example, the experiment by Hainmueller and Hopkins (2015) exploring the effect of country of origin on preference to be admitted as an immigrant in the U.S. had no option but to include a wide range of countries to explore the landscape of U.S. immigration. Instead of using countries, the authors could have used continents or other larger geographical units. However, it can be argued that biases against citizens of certain countries are not fully reflected in biases toward their respective continents and the other way around. In such contexts, the trade-off between the number of levels and statistical power is a crucial step in the design of conjoint experiments.

5.3 Type S and Type M errors

Traditional power analysis focuses on the probability that we correctly reject the null hypothesis when a specific alternative hypothesis is true. However, when an experiment does not reach adequate power (i.e. ≤ 0.80), the probability that the estimated effect sign is incorrect or the magnitude is biased increases substantially.

The results of the simulation highlight that, in addition to statistical power, Type S and Type M error rates are considerably important for conjoint experimental designs. In line with previous literature on the topic, Type S errors occur more frequently when experiments are underpowered. Results in Figure 6 show that the rate of Type S error spikes when both the sample used for the analysis ($n_{eff} = n \times t$) is small and when the true effect size is relatively low ($0.01 \leq \delta \leq 0.02$). On the contrary, when population

effects are larger ($\delta \geq 0.03$), Type S errors are infrequent.

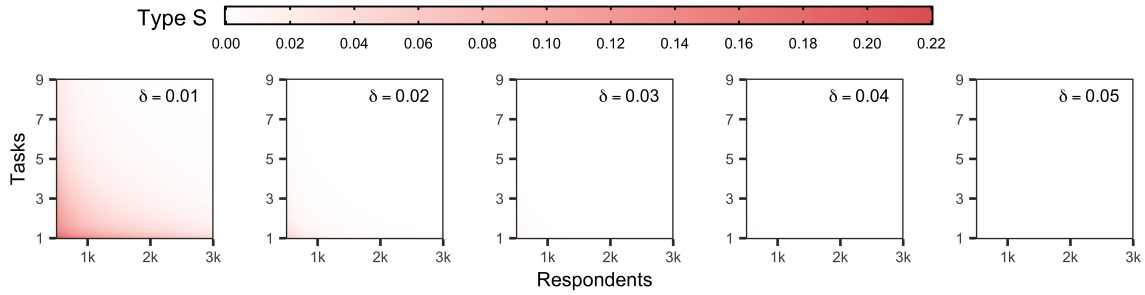

Figure 6: Type S error rates for different assumed AMCEs (δ). Shown values reflect a design with 2 levels ($l = 2$).

Similarly to statistical power, the Type S error rate is also affected by changing the number of levels in a conjoint design. Results in Figure 7 show that in terms of keeping Type S error at bay, an attribute with a large number of levels ($l \geq 15$) can be as demanding for sample size as searching for a small effect ($\delta \leq 0.02$). Assuming a large true effect size ($\delta = 0.05$), a common design with more than 15 levels and a relatively low sample carries a substantial risk (more than 10%) of concluding the estimated effect is correct when, in fact, has the opposite sign compared to the true population effect. Given the relatively small magnitudes of estimated effects and the large number of levels employed in the reviewed conjoint literature, this has an immediate impact on the ability to replicate some of the findings provided by conjoint experiments. Further, these false findings can also serve as statistical evidence against perfectly valid theories.

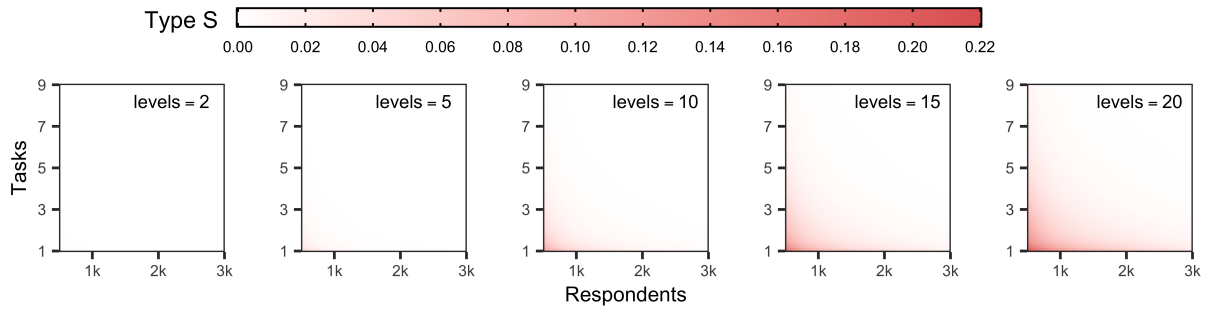

Figure 7: Effect of the number of levels (l) of the largest attribute on Type S error rates. Shown values assume a true AMCE (δ) = 0.05.

Another concern for conjoint experiments is the Type M error—exaggeration ratio.

Type M error is defined as the ratio of the absolute value of the estimated AMCE and the true effect size when the AMCE is found to be statistically significant. It is a measure of the expected systematic bias of the AMCE estimate, given the experimental design. Figure 8 reveals that Type M error is especially pronounced for small samples that find statistically significant effects mostly due to random fluctuations. The results show that conjoint experiments with small samples are likely to grossly overestimate small and potentially inconsequential effects. Specifically, an effect size of $\delta = 0.01$ estimated with a relatively small sample size ($n \leq 3000$ and $t \leq 3$) is, on average, likely to overestimate the true effect between 3- to 10-times (see Figure 8).

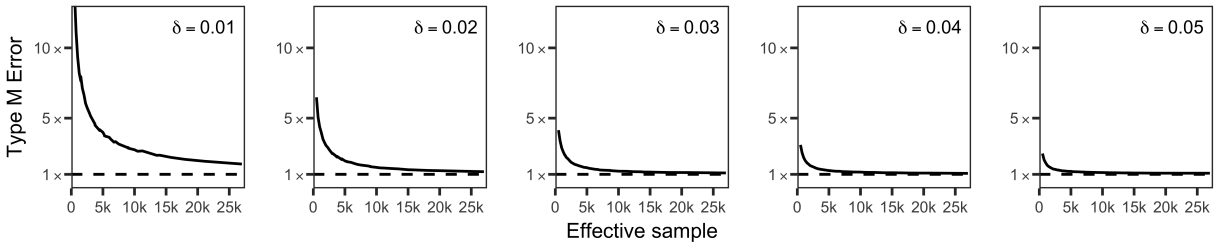

Figure 8: Type M error rates by sample size (n_{eff}) for different assumed AMCEs (δ).

5.4 Retrospective power analysis

A priori design considerations are not the only scenario where researchers incorporate information derived from past studies. Researchers can also explore whether statistically significant effects discovered in the literature change depending on the plausible size of the effect or in relation to particular design choices (Gelman and Carlin 2014, for an example see,). In this case, the question is whether the designs commonly employed in the conjoint literature are appropriate to find the expected, unbiased, and statistically significant effects. To this end, I performed a retrospective power analysis for the conjoint studies included in the literature review. Since the estimated effects found in the literature review are potentially biased, retrospective power has been calculated for each study based on a range of *assumed* true AMCEs ($0.01 \leq \delta \leq 0.15$). The emphasis on *assumed* is important here since it is virtually impossible to know the true effect sizes in the

population of reference.

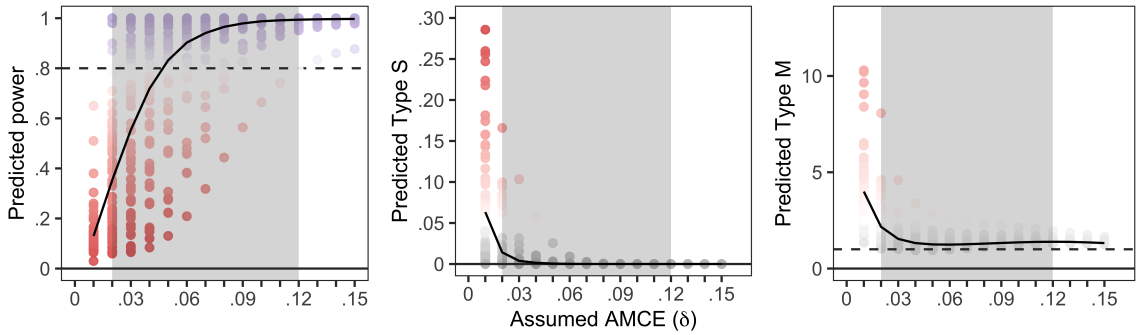

Figure 9: Retrospective statistical power, Type M and Type S error rates.

Figure 9 shows the calculated statistical power, Type S, and Type M error rates for the studies, collected in the literature review, across a range of assumed AMCEs. Each study is represented by a set of points—one point for each assumed AMCE. Results suggest that a substantial portion of the reviewed conjoint designs are underpowered for a range of assumed AMCEs. For true effect sizes that are equal to or smaller than 0.03 ($\delta \leq 0.03$), almost all of the designs employed in the reviewed conjoint literature are inadequate: on average, 89% of the conjoint studies under review would be considered underpowered. Among these, there is, on average, a 3% probability that a significant effect carries a wrong sign and the estimated effect is, on average, 2.5 times larger than the true effect. Passing to a true effect size between $0.03 \leq \delta \leq 0.06$, 34% of the conjoint studies included in the literature review are considered underpowered. In this case, the Type S error rate virtually drops to zero, while the average expected exaggeration factor is 1.2 times the true effect. For effect sizes that are between $0.06 \leq \delta \leq 0.09$, only 9% of the reviewed designs would be considered underpowered, with, on average, zero Type S error rate and an expected exaggeration factor of 1.3 times the true effect. Finally, for effect sizes $\delta \geq 0.09$, almost all reviewed studies seem well-powered, again with zero Type S error rate and an exaggeration rate of 1.3 times the true effect. In other words, the majority of used conjoint designs in the literature are well-powered to discover assumed AMCE of $\delta \geq 0.09$, where the overestimation of the magnitude of the effect is likely to be small and the direction of any statistically significant estimate is likely to have a correct

sign.

In the literature review, I reported that the median estimated AMCE is $\delta_{p_{50}} = 0.05$, ranging between $\delta_{p_{25}} = 0.025$ and $\delta_{p_{75}} = 0.12$ in its quartiles (grey area in Figure 9). Although it is likely that some of these estimates are biased due to low power, I take them as effects that have made it through peer review in the publication process, hence effects of this size are likely of interest to the research community. Assuming researchers want to discover effects of this size, the simulation shows that the currently used designs may prove insufficient: about one-third of the studies that I examined do not have sufficient power to discover effects of this size and are more prone to errors of sign and magnitude compared to designs recruiting larger samples or asking respondents for more trials.

6 Tools for applied research

6.1 Web application

One of the goals of this thesis is to provide guidance on how the number of experimental conditions, trials, and subjects impact the statistical power of conjoint experiments. Setting up a simulation like the one presented in this work is a rather complex task that requires extensive knowledge. Consequently, I designed a web application to calculate the power of simple conjoint designs based on the Hainmueller, Hopkins, and Yamamoto (2014) framework. The application uses the Shiny package (W. Chang et al. 2017) in the R programming language (R Core Team 2019). It requires no programming knowledge and can be used on any device with an internet connection or through the R Studio suite. This user-friendly application is available in having opened this link and its source code is on GitHub

The application has 4 main inputs that are located on the left panel of the app (see Figure 10).

- Respondents: Number of respondents that are going to answer the survey.

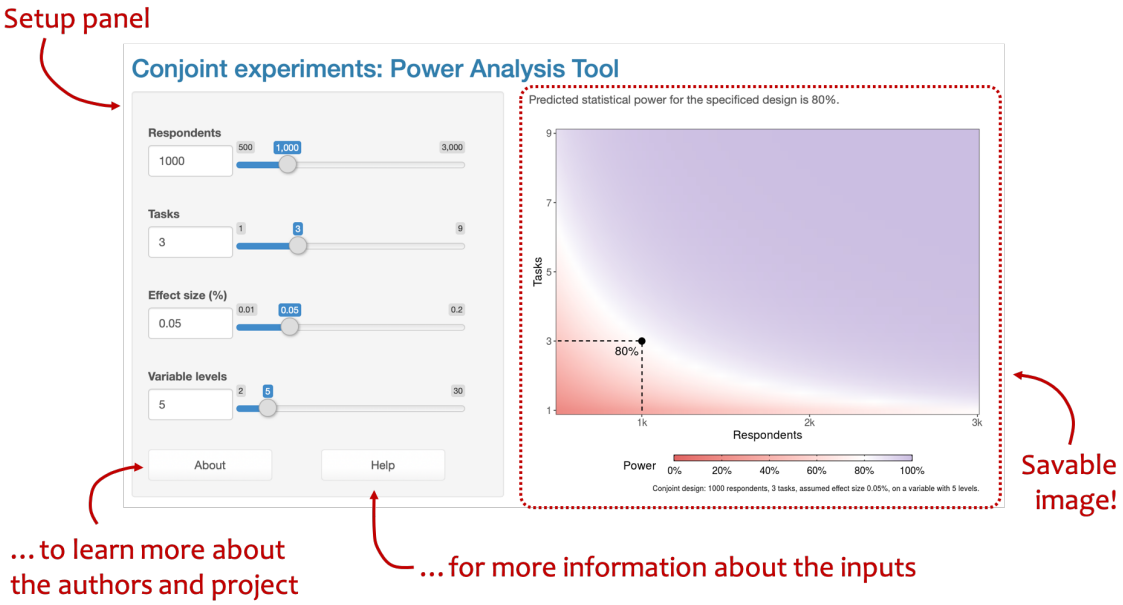

Figure 10: Shiny application for predicting the statistical power of conjoint experiments.

- Tasks: Number of tasks each respondent will receive (sometimes called trials or selection tasks).
- Tasks: Effect size: The expected effect size in %. This is the expected AMCE.
- Variable levels: The number of levels of your categorical variable — i.e. gender (male vs. female) has two categories.

The application produces graphs with the predicted statistical power, Type S and Type M errors (see infra for details on the predictions) based on the design input. The main goal of the application is to allow researchers to perform *a priori* power analysis. The application is particularly useful to evaluate the trade-off between different design aspects and adjust the design to achieve adequate power. For instance, let's assume that a researcher desires to discover a hypothesised AMCE of 5% using an $\alpha \leq 0.05$, but can only recruit a sample of 1000 respondents. Assuming that they want to have a 90% probability of discovering the hypothesised effect if the effect actually exists in the population, they could either increase the number of tasks per respondent or reduce the number of attribute levels to reach the desired statistical power. Another scenario of interest is the trade-off between the number of respondents and the number of tasks.

Let's assume that a researcher is interested in effect sizes of $\delta =\geq 0.05$ and can collect data from a sample between 500 and 1000 respondents. In this case, the application suggests that doubling the number of tasks each respondent completes is about as good as doubling the sample size.

While using the application is straightforward, two important caveats should be noted. First, power should be calculated for the attribute with the expected smallest AMCE. This ensures that all the attribute levels included in the design are properly powered, including the smallest one. Second, the experiment should be powered for the attribute with the highest number of levels. As mentioned before, power is affected by the number of levels of a given attribute due to the lower number of respondents *treated* within each level.

6.2 Predictive model

To translate the simulation results into the web application, I utilized a predictive model that uses the 4 main inputs located on the left panel of the application and returns the predicted statistical power, Type S, and Type M errors from the simulation. Lu, Qiu, and Deng (2019) proved that the power function and the probability of Type S and Type M error for a fixed significance level α is monotonic. Consequently, a linear model can be used to interpolate the simulation results. That is the estimation of the values that are between two known data points used in the simulation. The linear interpolation has been performed by fitting a generalised linear model using interactions and polynomials to predict new data points within the range of the simulation parameters. The usage of polynomials gives flexibility to the model and takes into account the exponential distribution observed in the trade-off between the various design elements.

One of the most important considerations in the interpolation between data points is the accuracy of the estimated output of the model. Specifically, it considers how precise the predicted interpolated points are compared to the observed data values generated by the simulation. Overall accuracy was determined by using the root-mean-square error (RMSE). The RMSE is the average distance of an observed data point from the fitted line

by the model. High RMSE values indicate that the model produced values that are further away from the regression line. On the contrary, a low RMSE indicates that the model provides good accuracy—that is, the distance between the observed data points and the predicted regression line is relatively small. The RMSE for the model of power (a logit model), Type S (a logit model) and Type M (an OLS with a logged dependent variable) errors are 0.023, 0.013, and 0.090, respectively. Figure 11 shows that the predicted data points for all simulation conditions are within reasonable margins of error: power is predicted in the majority of cases within ± 5%, Type S within ± 5%, and Type M within ± 10% of the observed value in the simulation. Although not perfect, I deemed the fit of this model satisfactory for the purpose of calculating the minimal required sample for conjoint experiments in the social sciences.

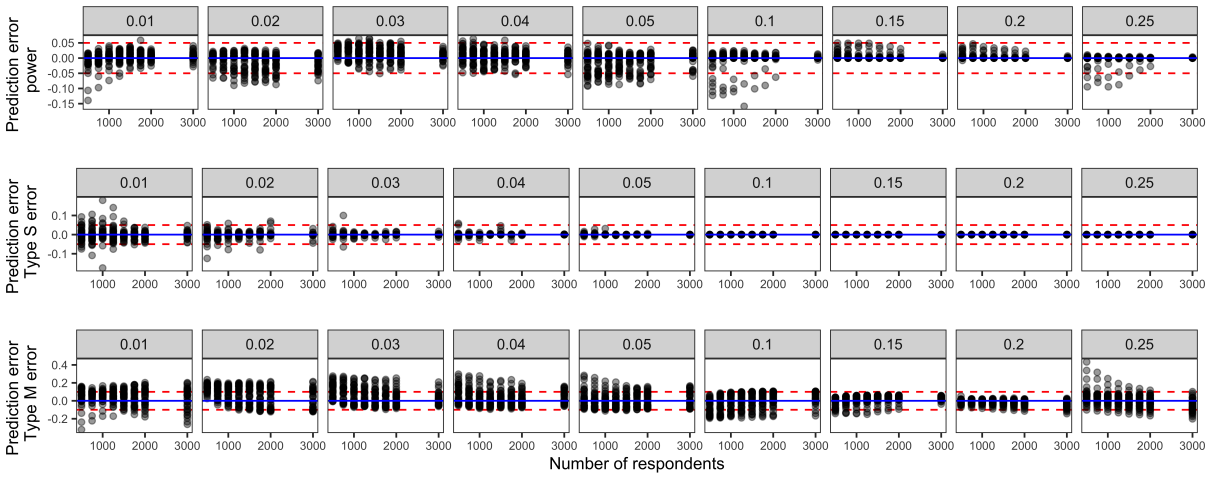

Figure 11: Error Rate predictive model for power, Type S and Type M

6.3 R function to simulate conjoint data

In addition to the web application, I the R code provided on GitHub can be used to set up a simulation based on the described data-generating process. This allows us to calculate power for more complex or less common designs that have not been included in the initial simulation. In its current developmental version, the function can be used to calculate minimum required sample size or power, Type S, and Type M error rates using (a) the number of attributes, (b) the number of levels per each attribute, (c) the hypothesised

(true) effect size, (d) the number of tasks, and (e) the number of respondents. The function also allows to change the treatment heterogeneity between respondents σ_k and thus to calculate power when respondent-level deviations u_{klj} are more pronounced.

I showcase how the function can be used to calculate power, Type S, and Type M error rates for a simple conjoint experiment with 5 attributes having 2 levels each where 1000 respondents are asked to complete 5 tasks. The treatment heterogeneity between respondents σ_k is fixed at 0.10.

```
# Number of attributes in the conjoint
n_attributes <- 5

# Number of levels per each attribute
# (numeric vector of length n_attributes)
n_levels = c(2, 2, 2, 2, 2)

# Hypothesized effect size (numeric vector of
# length sum(n_levels) - n_attributes)
simulation_coefs <- c(-0.04, 0.02, 0.00, -0.02, -0.05)

# Number of Respondents
sample_size <- c(1000)

# Number of tasks
num_tasks <- c(3)

# Treatment heterogeneity
sigma <- 0.10

# simulation runs
sim_runs <- 100

# seed for reproducibility
set.seed(999)
```

Next, I create a data frame to store the results of the simulation.

```
results <- data.frame(num_respondents=NA,
                      num_tasks=NA,
                      num_attrbs=NA,
                      num_lvls=NA,
                      true_coef=NA,
                      est_coef=NA,

                      est_se=NA,
                      est_se_robust=NA,

                      sig=NA,
                      sig_robust=NA,

                      in_ci95=NA,
                      in_ci95_robust=NA,

                      typeS=NA,
                      typeS_robust=NA,

                      typeM=NA,
                      typeM_robust=NA
)
```

After, I use the `generate_design()` function to generate the full factorial experimental design that is used to randomly sample the attribute levels for each respondent and experimental task.

```
sim_design <- generate_design(n_profiles = 2,
                              n_attributes = n_attributes,
                              n_levels = n_levels)
```

Finally, I use a loop function to run the simulation 100 times. Based on the specified design, the function generates a random sample of respondents (using `generate_sample()`), simulates the conjoint data (using `simulate_conjoint()`) and evaluates the model results

(using `evaluate.model()`).

```
for(s in 1:sim_runs){
  sample <- generate_sample(design = sim_design,
                            units = sample_size,
                            n_tasks = num_tasks)


  cj <- simulate_conjoint(data=sample,
                          coef = simulation_coefs,
                          sigma.u_k = sigma,
                          LOG = F)

  evaluate <- evaluate.model(cj,
                             y ~ as.factor(var_1) + as.factor(var_2) +
   as.factor(var_3) +
                               as.factor(var_4) + as.factor(var_5),
                             true_coefs = simulation_coefs)


  names(results) <- names(evaluate)
  results[s,] <- evaluate
}
```

After, I can calculate power, Type S, and Type M error rates taking the average of the 100 runs.

```
mean(results$sig_robust)
mean(results$typeS_robust, na.rm=T)
mean(results$typeM_robust, na.rm=T)
```

According to the results of the simulation, this particular conjoint design reaches a satisfactory level of power equal to .88 with no error of sign (Type S error) and an exaggeration ration of 1.13.

The R code contained in `CJpower_parallelsimulation.R` implements parallel computing to leverage on modern multi-core processors. The scripts also showcase how to simulta-

neously run the simulation for more than 1 design. Other aspects of the simulation (e.g., the significance level α) can be changed directly modifying the function contained in the script `CJpower_functions.R`. The R code and the functions showcased in this section are contained on the author's GitHub and can be found at this link.

7 Conclusions

This study showed that researchers need to be careful in designing conjoint experiments. I conducted numerous simulations to determine the appropriate sample size for various experimental designs commonly utilized in the conjoint literature political science research. In line with conventional understanding, statistical power grows with larger samples—be it by recruiting a larger number of respondents or adding another trial for each respondent. The simulations also show that the statistical power of a conjoint experiment decreases with a higher number of levels of the largest attribute. Nevertheless, contrary to widespread understanding, conjoint experiments require a substantial sample size to discover commonly sought small effects, especially for designs with high numbers of levels.

The results also show that underestimating the required sample can lead to biased and even opposite estimated effects. For each design included in the simulation, this study assessed the probability that the estimated parameter has an incorrect sign (Type S error) and how biased (exaggerated) it is compared to the true effect (Type M error). Results show that Type S and Type M errors are especially pronounced for experimental designs with relatively small sample sizes ($n_{eff} = n \times t \leq 3000$) or high numbers of attribute levels ($l \geq 10$) trying to estimate relatively small effects ($\delta \leq 0.03$). This indicates that for a wide range of commonly discovered AMCEs, the designs typically employed in political science are underpowered and thus are likely to result in biased estimates, both in terms of direction and magnitude. This finding should raise caution in the context of the replication crisis in psychology (Open Science Collaboration 2015). Even if conjoint experimental designs typically utilise samples much larger than experiments in psychology, the relatively high number of attributes-levels used in conjoint experiments can result in

designs with insufficient statistical power.

Generally, this study supports the idea that disregarding power considerations for experimental designs can have serious consequences for theory testing since it might lead researchers to pay attention to the wrong experimental factors or undermine relevant and valid theories. Consequently, designing conjoint experiments that maximise the chances of discovering the true effect and avoid inferential errors is important to prevent the publication of exaggerated results that are unlikely to be scientifically replicated. With this in mind, the simulation results have been used to estimate a predictive model that can be used by applied researchers to make sense of trade-offs associated with the expected number of respondents, trials, and levels. To that end, I have also released an open-source R code to enable researchers to investigate the trade-offs associated with statistical power in more complex scenarios. These include, for example, the presence of two or more subgroups of respondents, multi-stage sampling, or when individual-level heterogeneity is expected to be more pronounced.

All in all, this study demonstrates that power analysis is a useful tool in designing a conjoint study. That being said, the study presents several limitations. First, the results presented in the text are conditional on the match between the true and simulated data-generating models. This is true for every type of parametric simulation where a model is used to generate the data. Parametric simulations rely on a set of assumptions about the underlying data-generating mechanisms. In the context of this study, I assumed that the probability of selecting a profile is a linear combination of attribute levels shown in one profile (i.e., between-attribute comparison) *controlling* for the attribute levels shown in the other profile (i.e., within-attribute comparison). This assumption is rooted in past research on conjoint experiments that show that respondents engage in both types of cognitive processes when making a decision (Jenke et al. 2020; Meißner and Decker 2010). The insights obtained from the eye-tracking data are also backed up by recent research that uses a different estimator for conjoint data that specifically focus on the trade-off between the attribute level contained in one profile versus the attribute level

shown in the other profile (see, for instance, the recent article of Graham and Svolik 2020). Furthermore, the general validity of the proposed data-generating process is also confirmed by past research on factorial experiments (e.g., van Breukelen and Candel 2012; Jordan J. Louviere et al. 2000) and by the results obtained using the non-parametric approach proposed by Schuessler and Freitag (2020).

In spite of the assumed validity of the presented results, the data-generating process remains highly stylised with the absence of a proper validation procedure. Validating parametric simulations is not straightforward: it involves assessing the accuracy and reliability of the results by comparing them with real-world observations or other alternative data-generating mechanisms. Concerning the former, this can be achieved using an experiment where the true data-generating model is known. However, given the complex nature of conjoint experiments, I assume it is almost impossible to approximate the true data-generating process, even with the usage of sophisticated eye-tracking devices (Jenke et al. 2020). A viable alternative may involve drawing a set of samples from a study that is assumed to be well-powered and assessing how biased the estimate is as N varies. It is important to underline that, since the true data-generating process is still unknown, this procedure does not fully validate the assumed data-generating process. However, it would provide useful insights into whether the results using simulated data match real-world data.

Concerning the latter, additional parametric simulations can be used to investigate different data-generating mechanisms based on alternative decision-making mechanisms. Previous research has argued that respondents employ different types of decision-making strategies (e.g., weighted adding, satisficing strategies, boundedly rational) (Bettman, M. F. Luce, and Payne 1998). Thus, the data-generating process could be expanded to take into account that respondents may make choices according to different decision-making models. Subsequently, a sensitivity analysis could be performed by varying the parameters of the assumed behavioural model. This would allow assessing how the model results change if a sizeable percentage of the sample (e.g., $\geq 10\%$) diverges from the

assumed underlying model (i.e., both within- and between-attributes comparison) and uses a different type of behaviour model (e.g., only within-attribute comparison).

Third, compared to the previous approaches (e.g., Gall 2020), the used simulation covers a vast range of designs and can be easily modified to include more complex theoretical scenarios. However, it is limited in regard to two crucial aspects. In the simulation, the α level is fixed to ≤ 0.05 preventing researchers to investigate the trade-off between Type I and Type II errors when designing a conjoint experiment. The inability to manipulate the significance level α also implies that the simulation disregards Type I error inflation due to multiple comparisons. In the context of multi-dimensional like conjoint experiments, many different treatment effects are usually estimated. When conducting multiple comparisons, the likelihood of obtaining at least one significant result purely by chance increases. To overcome this issue, methods like the Bonferroni adjustment can be used to mitigate Type I errors (i.e., false positives) (Wright 1992). Practically, this can be performed by dividing the overall significance level (e.g., 0.05) by the number of comparisons. Further research should add these two important aspects when considering power analysis for conjoint experiments.

Fourth, the simulation also fixes the treatment heterogeneity between respondents $\sigma_k = 0.02$. Consequently, it is unable to test scenarios where respondent-level deviations u_{klj} are more pronounced. This could be relevant in two different theoretical scenarios where the usage of clustered standard errors may be required. The first one is when there is a substantial amount of treatment heterogeneity at the individual level that cannot be accounted for using observed covariates (i.e., subgroup analysis, see infra). In this case, respondents may employ different (unknown) decision-making mechanisms. If respondents are consistent across tasks and there is a high level of heterogeneity between respondents, this may impact the precision of the estimates and, thus, the power of the experiment. A second, related, scenario is when the conjoint design is relatively small (e.g., 2 attributes with 2 levels each), but the number of tasks is large (e.g., ≥ 20). Following Abadie et al. (2023), this is a situation where the within-individual correlation of treatment and

residuals is high and, thus, clustered errors are needed to retrieve unbiased estimates of the treatment effect. Intuitively, the reason is that each subsequent task will carry a lot of information since the attribute level displayed in task t will be rather similar compared to the one in task $t-1$. Further research should investigate these two theoretical scenarios and assess the minimum number of respondents needed to benefit from the cluster robust variance matrix estimator's asymptotic properties.

Last but not least, this study omits a substantial part of analysing conjoint experiments: interaction effects (Egami and Imai 2018). First, an interaction may be hypothesised between two experimental attributes in determining the probability of choosing a profile (for example, gender interacting with education in Figure 2). Second, the effect of an experimental attribute might differ based on some respondent's characteristics—for example, the effect of party affiliation in a conjoint profile being influenced by respondent's party identification (e.g., Kirkland and Coppock 2018). This type of interaction is usually called subgroup analysis and it is used to assess if treatment effects are significantly different across observed (Leeper, Hobolt, and Tilley 2019) or unobserved (Goplerud, Imai, and Pashley 2022) subgroups of respondents.

Although the standards for power analysis of interaction effects are ambiguous in the literature, some guidance is provided by Gall (2020) and (Schuessler and Freitag 2020). For both types of interaction effects, the minimum required sample size for a well-powered conjoint experiment rapidly increases. The functional relationship between interaction effects and power is partially revealed when we look at the relationship between the sample and the effect size. In the case of an interaction between two experimental attributes, it is common to find that an interaction alters the size, rather than the direction, of an existing attribute effect. In such cases, the required minimum sample size could double, triple, or even quadruple based on the effect of the moderator (on this point see results from, Brookes et al. 2004; Marshall 2007). When using subgroup analysis (interacting an attribute with respondents' characteristics or splitting the sample by respondent-varying covariates), the sample size decreases as a function of the number of groups included in the

analysis. For example, by splitting the sample by gender of the respondent (distributed as 50% female and 50% male), the sample used for estimating the effect within each subgroup is half of the original sample. Furthermore, when respondents belong to an unequally observed group (e.g., 70% female and 30% male respondents) the sample size used to estimate the effect for the male subgroup is substantially lower. This will be reflected in larger standard errors and subsequently lower power to obtain a significant result of a non-null effect. These examples show that the minimum required sample can increase drastically depending on both the size of the interaction effect and the number and size of the groups included in the analysis.

Beyond that, I suggest caution regarding new estimators that have been recently proposed. Though they tackle important limitations of the conjoint design, they often lack insight on minimal sample size requirements. For example, Ganter (2019) proves that the AMCE is not well suited to answer preference-related questions since it is influenced not only by the preference of the respondent on a given attribute but also by how the attributes are distributed in the population of interest (**@abramson_2019**). As such, the author proposes a new estimator—the Average Component Preference—that decomposes the AMCE into an effect of preference and effect of composition (how the attributes are distributed). This new quantity of interest then estimates the causal effect of an attribute X by looking only at profiles that differ in the levels of that attribute (only profiles where $X_{kp1} \neq X_{kp2}$). Although this approach allows for practical effect decomposition, it sacrifices a substantial part of the sample since two profiles are compared only if they have the same attribute level. In terms of statistical power, removing profiles that match on attribute levels leads to a substantial reduction in the number of observations used to estimate the treatment effects, and, thus, to an increase in the required sample size to achieve adequate power. By all means, we encourage future research into interaction effects and sample size considerations in conjoint experiments.

References

Abadie, Alberto et al. (2023). "When should you adjust standard errors for clustering?" In: *The Quarterly Journal of Economics* 138.1, pp. 1–35.

Adamowicz, W., J. Louviere, and M. Williams (1994). "Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities". In: *Journal of Environmental Economics and Management* 26.3, pp. 271–292. ISSN: 0095-0696. DOI: https://doi.org/10.1006/jeem.1994.1017. URL: https://www.sciencedirect.com/science/article/pii/S0095069684710175.

Adamowicz, Wiktor et al. (1998). "Stated preference approaches for measuring passive use values: choice experiments and contingent valuation". In: *American journal of agricultural economics* 80.1, pp. 64–75.

Arceneaux, Kevin (2005). "Using cluster randomized field experiments to study voting behavior". In: *The Annals of the American Academy of Political and Social Science* 601.1, pp. 169–179. DOI: 10.1177/0002716205277804.

Arnold, Benjamin F et al. (June 2011). "Simulation methods to estimate design power: an overview for applied research". In: *BMC Medical Research Methodology* 11.1. DOI: 10.1186/1471-2288-11-94. URL: https://doi.org/10.1186/1471-2288-11-94.

Astivia, Oscar L. Olvera, Anne Gadermann, and Martin Guhn (May 2019). "The relationship between statistical power and predictor distribution in multilevel logistic regression: a simulation-based approach". In: *BMC Medical Research Methodology* 19.1. DOI: 10.1186/s12874-019-0742-8. URL: https://doi.org/10.1186/s12874-019-0742-8.

Baker, Monya (2016). "Reproducibility crisis". In: *Nature* 533.26, pp. 353–66.

Bansak, Kirk et al. (2019). "Beyond the breaking point? Survey satisficing in conjoint experiments". en. In: *Political Science Research and Methods*. Publisher: Cambridge University Press, pp. 1–19. ISSN: 2049-8470, 2049-8489. DOI: 10.1017/psrm.2019.13. URL: http://www.cambridge.org/core/journals/political-science-research-

and-methods/article/beyond-the-breaking-point-survey-satisficing-in-conjoint-experiments/47EBFC9D8CE2CA88D6C01870192F8956 (visited on 12/29/2020).

Bechtel, Michael M and Roman Liesch (July 9, 2020). "Reforms and Redistribution: Disentangling the Egoistic and Sociotropic Origins of Voter Preferences". In: *Public Opinion Quarterly* 84.1, pp. 1–23. ISSN: 0033-362X, 1537-5331. DOI: 10.1093/poq/nfaa006.

Bettman, James R, Mary Frances Luce, and John W Payne (1998). "Constructive consumer choice processes". In: *Journal of consumer research* 25.3, pp. 187–217.

Boxall, Peter C. et al. (1996). "A comparison of stated preference methods for environmental valuation". In: *Ecological Economics* 18.3, pp. 243–253. ISSN: 0921-8009. DOI: https://doi.org/10.1016/0921-8009(96)00039-0. URL: https://www.sciencedirect.com/science/article/pii/0921800996000390.

Breur, Tom (2016). "Statistical Power Analysis and the contemporary "crisis" in social sciences". In: *Journal of Marketing Analytics* 4.2-3, pp. 61–65.

Brookes, Sara T et al. (2004). "Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test". In: *Journal of clinical epidemiology* 57.3, pp. 229–236.

Carson, Richard T et al. (1994). "Experimental analysis of choice". In: *Marketing letters* 5, pp. 351–367.

Chang, Mark and Shein-Chung Chow (2006). "Power and Sample Size for Dose Response Studies". In: *Dose Finding in Drug Development*. Ed. by Naitee Ting. Statistics for Biology and Health. New York, NY: Springer, pp. 220–241. ISBN: 978-0-387-33706-7. DOI: 10.1007/0-387-33706-7_14.

Chang, Winston et al. (2017). "Shiny: Web Application Framework for R". In: *R package version* 1.5.

Cohen, Jacob (1962). "The statistical power of abnormal-social psychological research: a review." In: *The Journal of Abnormal and Social Psychology* 65.3, p. 145. DOI: 10.1037/h0045186.

— (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cox, David Roxbee (1958). "Planning of experiments." In.

de la Cuesta, Brandon, Naoki Egami, and Kosuke Imai (Jan. 2022). "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution". In: *Political Analysis* 30.1, pp. 19–45. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2020.40. (Visited on 03/17/2022).

Druckman, James N. et al., eds. (2011). *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge Univ. Press. ISBN: 978-0-521-19212-5 978-0-521-17455-8.

Duch, Raymond M. et al. (2020). "Nativist Policy: The Comparative Effects of Trumpian Politics on Migration Decisions". In: *Political Science Research and Methods*, pp. 1–17. ISSN: 2049-8470, 2049-8489. DOI: 10.1017/psrm.2020.33.

Dziak, John J., Inbal Nahum-Shani, and Linda M. Collins (2012). "Multilevel Factorial Experiments for Developing Behavioral Interventions: Power, Sample Size, and Resource Considerations". In: *Psychological Methods* 17.2, pp. 153–175. ISSN: 1939-1463(Electronic),1082-989X(Print). DOI: 10.1037/a0026972.

Egami, Naoki and Kosuke Imai (Aug. 2018). "Causal Interaction in Factorial Experiments: Application to Conjoint Analysis". In: *Journal of the American Statistical Association* 114.526, pp. 529–540. DOI: 10.1080/01621459.2018.1476246. URL: https://doi.org/10.1080/01621459.2018.1476246.

Franchino, Fabio and Francesco Zucchini (May 2015). "Voting in a Multi-dimensional Space: A Conjoint Analysis Employing Valence and Ideology Attributes of Candidates". In: *Political Science Research and Methods* 3.02, pp. 221–241. ISSN: 2049-8470, 2049-8489. DOI: 10/gf8qgs. (Visited on 08/26/2017).

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk (2007). "The Logic of the Survey Experiment Reexamined". In: *Political Analysis* 15.1, pp. 1–20. DOI: 10.1093/pan/mpl008. URL: https://doi.org/10.1093/pan/mpl008.

Gall, Brett J (Oct. 2020). *Simulation-based power calculations for conjoint experiments*. DOI: 10.31219/osf.io/bv6ug. URL: osf.io/bv6ug.

Ganter, Flavien (Dec. 2019). *Identification of Preferences in Forced-Choice Conjoint Experiments: Reassessing the Quantity of Interest*. DOI: 10.31235/osf.io/e638u. URL: osf.io/preprints/socarxiv/e638u.

Gelman, Andrew and John Carlin (2014). "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors". In: *Perspectives on Psychological Science* 9.6, pp. 641–651. DOI: 10.1177/1745691614551642.

Goggin, Stephen N., John A. Henderson, and Alexander G. Theodoridis (Jan. 2019). "What Goes with Red and Blue? Mapping Partisan and Ideological Associations in the Minds of Voters". In: *Political Behavior*. ISSN: 1573-6687. DOI: 10/gmqpbx. (Visited on 02/05/2020).

Goplerud, Max, Kosuke Imai, and Nicole E. Pashley (Jan. 4, 2022). *Estimating Heterogeneous Causal Effects of High-Dimensional Treatments: Application to Conjoint Analysis*. arXiv: 2201.01357 [stat]. URL: http://arxiv.org/abs/2201.01357 (visited on 03/17/2022).

Graham, Matthew H and Milan W Svolik (2020). "Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States". In: *American Political Science Review* 114.2, pp. 392–409.

Green, Paul E and Vithala R Rao (1971). "Conjoint measurement-for quantifying judgmental data". In: *Journal of Marketing research* 8.3, pp. 355–363.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto (Feb. 2015). "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior". In: *Proceedings of the National Academy of Sciences* 112.8, pp. 2395–2400. ISSN: 0027-8424, 1091-6490. DOI: 10/f63mkr. (Visited on 05/23/2019).

Hainmueller, Jens and Daniel J. Hopkins (2015). "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants". In: *American Journal of Political Science* 59.3, pp. 529–548. ISSN: 1540-5907. DOI: 10.1111/ajps.12138.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto (2014). "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference

Experiments". In: *Political Analysis* 22.1, pp. 1–30. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mpt024.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960. ISSN: 0162-1459. DOI: 10/fjf6jb. JSTOR: 2289064. (Visited on 02/14/2020).

Huff, Connor and Joshua D. Kertzer (2018). "How the Public Defines Terrorism". In: *American Journal of Political Science* 62.1, pp. 55–71. ISSN: 1540-5907. DOI: 10/gcw73g. (Visited on 09/10/2020).

Jasso, Guillermina and Peter H Rossi (1977). "Distributive justice and earned income". In: *American Sociological Review*, pp. 639–651.

Jenke, Libby et al. (2020). "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments". In: *Political Analysis*, pp. 1–27. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2020.11.

Johnson, F Reed, Melissa Ruby Banzhaf, and William H Desvousges (2000). "Willingness to pay for improved respiratory and cardiovascular health: a multiple-format, stated-preference approach". In: *Health Economics* 9.4, pp. 295–317.

Kertzer, Joshua D, Jonathan Renshon, and Keren Yarhi-Milo (June 13, 2019). "How Do Observers Assess Resolve?" In: *British Journal of Political Science*, pp. 1–23. ISSN: 0007-1234, 1469-2112. DOI: 10.1017/S0007123418000595.

Khanna, Kabir (2019). *What traits are Democrats prioritizing in 2020 candidates?* URL: https://www.cbsnews.com/news/democratic-voters-hungry-for-women-and-people-of-color-in-2020-nomination/.

Kirkland, Patricia A. and Alexander Coppock (Sept. 1, 2018). "Candidate Choice Without Party Labels:" in: *Political Behavior* 40.3, pp. 571–591. ISSN: 1573-6687. DOI: 10.1007/s11109-017-9414-8.

Knudsen, Erik and Mikael Poul Johannesson (Sept. 13, 2018). "Beyond the Limits of Survey Experiments: How Conjoint Designs Advance Causal Inference in Political

Communication Research". In: *Political Communication* 0.0, pp. 1–13. ISSN: 1058-4609. DOI: 10.1080/10584609.2018.1493009.

Leeper, Thomas J., Sara B. Hobolt, and James Tilley (Aug. 7, 2019). "Measuring Subgroup Preferences in Conjoint Experiments". In: *Political Analysis*, pp. 1–15. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2019.30.

Loewen, Peter John, Daniel Rubenson, and Arthur Spirling (2012). "Testing the power of arguments in referendums: A Bradley–Terry approach". In: *Electoral Studies* 31.1. Special Symposium: Germany's Federal Election September 2009, pp. 212–221. ISSN: 0261-3794. DOI: https://doi.org/10.1016/j.electstud.2011.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0261379411000953.

Louviere, Jordan J. et al. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge University Press. DOI: 10.1017/CBO9780511753831.

Louviere, Jordan. J. (2015). "Measurement Theory: Conjoint Analysis Applications". In: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 865–867. DOI: 10.1016/b978-0-08-097086-8.43060-6. URL: https://doi.org/10.1016/b978-0-08-097086-8.43060-6.

Lu, Jiannan, Yixuan Qiu, and Alex Deng (2019). "A note on Type S/M errors in hypothesis testing". In: *British Journal of Mathematical and Statistical Psychology* 72.1, pp. 1–17. DOI: 10.1111/bmsp.12132.

Luce, R Duncan and John W Tukey (1964). "Simultaneous conjoint measurement: A new type of fundamental measurement". In: *Journal of mathematical psychology* 1.1, pp. 1–27.

Lukac, Martin and Alberto Stefanelli (2020). URL: https://mblukac.shinyapps.io/conjoints-power-shiny/.

Mackenzie, John (1993). "A Comparison of Contingent Preference Models". In: *American Journal of Agricultural Economics* 75.3, pp. 593–603. ISSN: 00029092, 14678276. URL: http://www.jstor.org/stable/1243566 (visited on 08/10/2023).

Marshall, Stephen W (2007). "Power for tests of interaction: effect of raising the Type I error rate". In: *Epidemiologic Perspectives & Innovations* 4.1, pp. 1–7.

Martin, Lucy (2019). "All Sins Are Not Created Equal: The Factors That Drive Perceptions of Corruption Severity". In: *Journal of Experimental Political Science*, pp. 1–11. ISSN: 2052-2630, 2052-2649. DOI: 10/gmqpdj. (Visited on 09/10/2020).

Maxwell, Scott E. (2004). "The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies." In: *Psychological Methods* 9.2, pp. 147–163. DOI: 10.1037/1082-989x.9.2.147.

Meißner, Martin and Reinhold Decker (Sept. 1, 2010). "Eye-Tracking Information Processing in Choice-Based Conjoint Analysis". In: *International Journal of Market Research* 52.5, pp. 593–612. ISSN: 1470-7853. DOI: 10/dfjnjz. URL: https://doi.org/10.2501/S147078531020151X (visited on 01/16/2022).

Moerbeek, Mirjam (Jan. 1, 2011). "The Effects of the Number of Cohorts, Degree of Overlap Among Cohorts, and Frequency of Observation on Power in Accelerated Longitudinal Designs". In: *Methodology* 7.1, pp. 11–24. ISSN: 1614-1881. DOI: 10.1027/1614-2241/a000019.

Neyman, Jerzy (1923). "On the application of probability theory to agricultural experiments. Essay on principles". In: *Ann. Agricultural Sciences*, pp. 1–51.

Open Science Collaboration (Aug. 2015). "Estimating the reproducibility of psychological science". In: *Science* 349.6251, aac4716–aac4716. DOI: 10.1126/science.aac4716. URL: https://doi.org/10.1126/science.aac4716.

Orme, Bryan (1998). "Sample Size Issues for Conjoint Analysis Studies". In: *Sawthooth Software Research paper Series. Squim, WA, USA: Sawthooth Software Inc.* URL: https://pdfs.semanticscholar.org/a696/a7f303fba91009aaaf1fb2db8c730f39ea05.pdf (visited on 03/17/2017).

Peterson, Erik and Gabor Simonovits (Aug. 22, 2018). "The Electoral Consequences of Issue Frames". In: *The Journal of Politics* 80.4, pp. 1283–1296. ISSN: 0022-3816. DOI: 10.1086/698886.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Rose, John M. and Michiel C. J. Bliemer (Sept. 2013). "Sample size requirements for stated choice experiments". en. In: *Transportation* 40.5, pp. 1021–1041. ISSN: 1572-9435. DOI: 10.1007/s11116-013-9451-z. URL: https://doi.org/10.1007/s11116-013-9451-z (visited on 09/29/2020).

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5, p. 688.

— (2005). "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469, pp. 322–331.

Ryan, Mandy and Shelley Farrar (2000). "Using conjoint analysis to elicit preferences for health care". In: *Bmj* 320.7248, pp. 1530–1533.

Schuessler, Julian and Markus Freitag (Dec. 2020). *Power Analysis for Conjoint Experiments*. DOI: 10.31235/osf.io/9yuhp. URL: osf.io/preprints/socarxiv/9yuhp.

Stefanelli, Alberto and Martin Lukac (Nov. 2020). *Subjects, Trials, and Levels: Statistical Power in Conjoint Experiments*. DOI: 10.31235/osf.io/spkcy. URL: osf.io/preprints/socarxiv/spkcy.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth (Aug. 2018). "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics". In: *American Political Science Review* 112.3, pp. 525–541. ISSN: 0003-0554, 1537-5943. DOI: 10/gdwd55. (Visited on 02/05/2020).

Train, Kenneth E. (2009). *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge University Press. DOI: 10.1017/CBO9780511805271.

Tversky, Amos (1967). "Additivity, utility, and subjective probability". In: *Journal of Mathematical psychology* 4.2, pp. 175–201.

Van Breukelen, Gerard J. P. and Math J. J. M. Candel (Nov. 1, 2012). "Calculating Sample Sizes for Cluster Randomized Trials: We Can Keep It Simple and Efficient!"

In: *Journal of Clinical Epidemiology* 65.11, pp. 1212–1218. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2012.06.002.

Wallander, Lisa (2009). "25 years of factorial surveys in sociology: A review". In: *Social Science Research* 38.3, pp. 505–520. ISSN: 0049-089X. DOI: https://doi.org/10.1016/j.ssresearch.2009.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0049089X09000192.

Ward, Dalston G. (Feb. 2019). "Public Attitudes toward Young Immigrant Men". In: *American Political Science Review* 113.1, pp. 264–269. ISSN: 0003-0554, 1537-5943. DOI: 10/gft5n7. (Visited on 07/08/2021).

Wright, S Paul (1992). "Adjusted p-values for simultaneous inference". In: *Biometrics*, pp. 1005–1013.

FACULTY OF SCIENCE
CELESTIJNENLAAN 200I
3001 HEVERLEE, BELGIË
tel. +32 16321401
www.kuleuven.be